

Signals and Communication Technology

Shanzhi Chen

Yan Shi

Bo Hu

Ming Ai

Mobility Management

Principle, Technology and Applications

 Springer

Signals and Communication Technology

More information about this series at <http://www.springer.com/series/4748>

Shanzhi Chen · Yan Shi · Bo Hu
Ming Ai

Mobility Management

Principle, Technology and Applications



Springer

Shanzhi Chen
China Academy of Telecommunication
Technology
Beijing
China

Bo Hu
Beijing University of Posts
and Telecommunications
Beijing
China

Yan Shi
Beijing University of Posts
and Telecommunications
Beijing
China

Ming Ai
China Academy of Telecommunication
Technology
Beijing
China

ISSN 1860-4862 ISSN 1860-4870 (electronic)
Signals and Communication Technology
ISBN 978-3-662-52724-5 ISBN 978-3-662-52726-9 (eBook)
DOI 10.1007/978-3-662-52726-9

Library of Congress Control Number: 2016941286

© Springer-Verlag Berlin Heidelberg 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer-Verlag GmbH Berlin Heidelberg

*To my parents, my wife Yali Gao and my son
Bob Junzhe Chen with loves and gratitude*

—Shanzhi Chen

*To my parents and family for their loves, and
to my teachers for their guidance*

—Yan Shi

To my parents and my wife Lin Si

—Bo Hu

To my parents, my wife and my son

—Ming Ai

Preface

Mobility management is the technology aiming to provide communication and service continuity for mobile objects (i.e., users, machines, and services). It originated from the cellular system as one feature in the specific network and developed continually as a general capability of networks in the following network evolution stages such as Internet, mobile Internet, non-infrastructure network, and Internet of Things (IoT). When prospecting the future mobility-driven network (MDN), it is envisioned that mobility plays an increasing role in communications and services, including mobility of humans in traditional mobile communications, mobility of machines in machine-to-machine (M2M) communications, mobility of contents in data-centric networks, mobility of service providers in new emerging applications, and mobility of computing resources in mobile cloud computing. The mobility as the inherent capability is used for implementing communications, networking, and applications. The mobility supporting will be one of the focuses going through the future network architecture developments and the essential technology research.

This book is based upon years of academic and industrial experiences in mobility management of the authors. It is the first book treating mobility management as an individual technology with systematic and insight view. The text has been prepared to provide fundamental concepts of technical principles and design philosophies yet avoiding discussions in protocol details.

The mobility management reference models proposed by authors are used in the book as the main track, offering multi-dimensional investigation and horizontal comparison of typical technologies. The model is used to explain profound theories in simple ways and simple languages, making it an excellent reference for the researchers, engineers, and students working in related areas.

This book covers major principles, technologies, and applications of mobility management technology. Chapters of the book are structured into three main parts.

Part I: Fundamental Concepts and Principles (Chaps. 1–4)

Chapter 1 provides an overview of the background, evolution, definitions, and classifications of mobility management, with brief introduction to standardization activities.

Chapter 2 presents reference models of the mobility management technology, including the protocol reference model identifying the critical control functions and network reference model abstracting the function entities. Typical mobility management technologies are analyzed accordingly.

Chapter 3 provides extensive discussions on technologies supporting different mobility objectives and highlights the technical principles rather than protocol details. This chapter is particularly useful for quickly understanding these technologies.

Chapter 4 discusses the applications of some mathematical theories in mobility management researches including mobility modeling, performance analysis, and algorithm design.

Part II: Classical Technologies and Applications (Chaps. 5–8)

Chapter 5 introduces the mobility management technology in the cellular communication system as the classical technology at the data link layer. Major network entities and control functions are abstracted and analyzed according to the reference models introduced in this book, especially the design principles in location management and handover control.

Chapter 6 selects Mobile IP (MIP) as the typical mobility management technology at the network layer. Extensive discussions on critical control functions in MIP are offered based on the reference models introduced in the book. Network mobility support capability of MIP is also discussed.

Chapter 7 takes the mobile Stream Control Transmission Protocol (mSCTP) as a typical protocol providing mobility support at the transport layer. Based on the introduction to existing schemes with explanations using their technical principles, mSCTP is introduced with in-depth analyses on the critical control functions.

Chapter 8 introduces the mobility support technology at the application layer, where Session Initiation Protocol (SIP) is taken as the typical protocol example. The control functions and its capabilities to support various mobility objectives are presented.

Part III: Future Requirements, Challenges, and Directions (Chap. 9)

Chapter 9 starts with analyses of the technical requirements of specific scenarios. Based on those requirements, the methodology-related challenges are derived and the research trends are presented.

Beijing, China
June 2016

Shanzhi Chen
Yan Shi
Bo Hu
Ming Ai

Acknowledgements

The authors wish to express sincere gratitude to several industry reviewers at Datang Telecom Technology & Industry Group, who offered helpful suggestions based on their rich research results and engineering experiences. We benefit greatly from valuable suggestions provided by Li Xu, Jian Zhao, Hucheng Wang, and Fangli Xu for mobility management technologies at data link layer (Chap. 5). The authors sincerely appreciate Yanfei Zhou and Yunjing Hou for their valuable inputs to Chap. 6.

We are also pleased to acknowledge the helps from graduate students of BUPT, who have helped us greatly in preparing related materials. Zhenxiang Gao, Changkai Lu, Li Li, Qingyun Xin, Xiaojuan Zhou, Xuemei Zhu, and Bin Chen surveyed the papers about applying mathematical theories in mobility management (Chap. 4). Yao Zhang, Huan Wang, and Ning Li prepared materials for mobility management technologies at the data link layer (Chap. 5) and the network layer (Chap. 6). Jianli Pan collected references for mobility management technologies at the application layer (Chap. 8).

Special thanks go to Prof. Jinhu Lyu and his doctoral student Shaolin Tan at the Institute of Mathematics and Systems Science in Chinese Academy of Sciences for their comprehensive review and valuable suggestions about applying mathematical theories in mobility management (Chap. 4).

Finally, it is a pleasure to acknowledge Lu Yang, Huijie Guo and Springer, who helped us to bring this book to you.

Contents

Part I Fundamental Concepts and Principles

1 Introduction	3
1.1 Background and Evolution	3
1.1.1 History	3
1.1.2 Research Background	7
1.1.3 Evolution Directions	8
1.2 Definitions	10
1.2.1 Mobility	10
1.2.2 Classification	11
1.2.3 Related Terminologies	15
1.3 Standardization Activities	17
1.3.1 3GPP	17
1.3.2 IETF	19
1.3.3 ITU	21
1.3.4 IEEE	22
1.3.5 ETSI	23
1.3.6 WFA	24
1.3.7 WBA	24
1.4 Motivation and Outline	24
References	25
2 Mobility Management Reference Models	29
2.1 Protocol Reference Model	29
2.1.1 Data Plane	29
2.1.2 Control Plane	33
2.1.3 Management Plane	34
2.2 Network Reference Model	35
2.2.1 Mobility Management Server	36
2.2.2 Access Point	37
2.2.3 Domain Mobility Controller	38
2.2.4 Mobility Management Protocol	38

2.3	Critical Control Functions in Mobility Management	39
2.3.1	Security Mechanism	39
2.3.2	Location Management	40
2.3.3	Handover Control	45
2.3.4	Interoperability Control	52
2.4	Analysis of Typical Mobility Management Technologies	56
2.4.1	Data Link Layer: Mobility Management in Cellular Systems	56
2.4.2	Network Layer: MIP	57
2.4.3	Transport Layer: MSCTP	59
2.4.4	Application Layer: SIP	59
	References	61
3	Mobility Objects and Technologies	63
3.1	Terminal Mobility	63
3.1.1	Host-Based Mobility Versus Network-Based Mobility	63
3.1.2	Horizontal Handover Versus Vertical Handover	66
3.1.3	Traditional Mobility Management Versus ID/Locator Separation	73
3.2	Personal Mobility	74
3.2.1	Mobile Agent-Based Schemes	75
3.2.2	Unified Mobility Management Database-Based Schemes	75
3.3	Service Mobility	77
3.3.1	Service Mobility at the Terminal Side	77
3.3.2	Service Mobility at the Provider Side	78
3.4	Network Mobility	79
3.4.1	Network Mobility Scenarios	79
3.4.2	Technical Solutions	80
3.4.3	Routing Optimization in Network Mobility	82
3.4.4	Applications of Network Mobility	83
	References	86
4	Application of Mathematical Theories in Mobility Management	91
4.1	Overview	91
4.2	Probability and Statistics Theory	91
4.2.1	Using Classical Distributions as Assumptions	92
4.2.2	Defining the Probabilities of Classical Events for Analysis	93
4.2.3	Use of the Bayes' Multi-criteria Decision Making for Handover Evaluation	95
4.3	Stochastic Processes	95
4.3.1	Modeling of Incoming Call Arrival	96
4.3.2	Mobility Modeling and Prediction	96
4.3.3	Location Management	100

4.4	Markov Process.	101
4.4.1	Performance Evaluation	101
4.4.2	Vertical Handover Decision	103
4.5	Queuing Theory	105
4.5.1	Queuing Strategies in Channel Allocation.	105
4.5.2	Delay Performance Analysis.	107
4.6	Multi-criteria Decision Making	108
4.6.1	Simple Weighted Sum-Based Solutions	108
4.6.2	Decision Function-Based Solutions	109
4.6.3	AHP-Based Solutions	109
4.6.4	TOPSIS-Based Solutions	110
4.7	Fuzzy Theory	110
4.7.1	Handover Decision	111
4.7.2	Location Management	113
	References	113

Part II Classical Technologies and Applications

5	Mobility Management at Link Layer	121
5.1	Overview of Typical Wireless Systems.	121
5.1.1	Cellular System	121
5.1.2	WLAN	126
5.1.3	Satellite Communication Systems	127
5.2	Overview of Network Architecture and Mobility Management in Cellular Systems	128
5.2.1	Network Architecture Evolution of Cellular Systems	128
5.2.2	Mobility Management Models of GSM/UMTS/LTE Systems	137
5.3	Mobility Management Within Radio Access Networks.	143
5.3.1	Mobility Management in the Idle Mode	143
5.3.2	Mobility Management in the Connected Mode	145
5.4	Location Management from Core Network Aspect.	151
5.4.1	General Ideas of Location Management	151
5.4.2	Idle-Mode Signaling Reduction.	157
5.4.3	Typical Location Management Procedures	158
5.5	Handover Control from Core Network Aspects	164
5.5.1	Handover Control Within CS Domain	164
5.5.2	Handover Control Within PS Domain	165
5.5.3	Handover Control for CSFB and SRVCC.	175
5.6	Security Mechanism.	179
5.6.1	Security Mechanisms in GSM, UMTS, and LTE.	179
5.6.2	Security Mechanisms and Mobility Management Interactions.	185

5.7	Interworking and Convergence Between WLAN and Cellular Networks	188
5.7.1	The Loose Coupling and Tight Coupling Interworking Architectures	189
5.7.2	I-WLAN Interworking Architecture	189
5.7.3	EPC-Centric Interworking Architecture	194
	References	198
6	Mobility Management at Network Layer	201
6.1	Overview	201
6.2	Mobility Management Solutions at Network Layer	202
6.2.1	Host-Based Mobility Management for Host Mobility	202
6.2.2	Network-Based Mobility Management for Host Mobility	211
6.2.3	Network Mobility	214
6.2.4	IP Flow Mobility	219
6.3	Key Control Technologies for Mobility Management at Network Layer	220
6.3.1	Overview	220
6.3.2	Location Management and Its Optimizations.	221
6.3.3	Handover Control and Its Optimization	227
6.3.4	Security Mechanism	231
	References	238
7	Mobility Management at Transport Layer	241
7.1	Overview	241
7.2	Transport-Layer Mobility Support Solutions	242
7.2.1	Gateway-Based Mobility Scheme	242
7.2.2	Connection Migration Protocol	244
7.2.3	Transport-Layer Handover Protocol	245
7.2.4	Complete Mobility Management Schemes	246
7.2.5	Comparison	247
7.3	Typical Protocol: mSCTP	248
7.3.1	Overview	248
7.3.2	Location Management	253
7.3.3	Handover Control	255
7.3.4	Security Mechanisms	260
	References	263
8	Mobility Management at Application Layer	265
8.1	Overview	265
8.2	Typical Protocol: SIP	267
8.2.1	Overview	267
8.2.2	Location Management	268
8.2.3	Handover Control	271
8.2.4	Security Mechanism	275

8.3	High-Level Mobility Support of SIP	277
8.3.1	Personal Mobility Support of SIP	277
8.3.2	Service Mobility Support of SIP	279
8.4	Applications of Mobility Support of SIP.	279
8.4.1	SIP-Based Mobility Support in Vehicular Networks.	279
8.4.2	SIP-Based Mobility Support in Satellite Communication Networks	280
	References	281

Part III Future Requirements, Challenges and Directions

9	Requirements, Challenges, and Trends	285
9.1	Scenario-Specific Requirements	285
9.1.1	The Application Perspective	285
9.1.2	The Terminal Perspective	288
9.1.3	The Network Perspective	291
9.2	Methodology-Related Challenges.	295
9.2.1	Identifiers and Addressing in Future Internet.	295
9.2.2	Emerging Radio Access Network Architectures.	297
9.2.3	Mobility Management Function Structures: Centralized Versus Distributed	298
9.2.4	Mobility Patterns.	300
9.3	Solution Trends.	301
9.3.1	Design ID–Locator Architecture for Mobility	301
9.3.2	Mobility Management Enhancements for Emerging Radio Access Network Architectures	302
9.3.3	Distributed Dynamic Mobility Management	302
9.3.4	Modeling of Mobility Patterns and Mobility Pattern-Based Mobility Management Solutions	303
	References	305
	Index	307

About the Authors

Shanzhi Chen received his Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT), China, in 1997. He achieved his M.S. (1994) from China Academy of Posts and Telecommunications (CAPT) and his B.E. (1991) from Xidian University, china. He is currently the Chief Technology Officer (CTO) for both Datang Telecom Technology & Industry Group and for China Academy of Telecommunication Technology (CATT). He is also the director of State Key Laboratory of Wireless Mobile Communications and the board member of Semiconductor Manufacturing International Corporation (SMIC). He is a Guest Professor for both BUPT and Beijing Institute of Technology (BIT).

Dr. Chen has more than 25 years of experience in the broadband communications and the wireless mobile communications, in both industry and academia. He was a visiting researcher at the Alcatel Bell research Center in Antwerp, Belgium, in 1996. He joined Datang Telecom Technology & Industry Group in 1994 and has been serving as CTO since 2008. He devoted his researches and developments to TD-LTE-advanced 4G since 2004 and made tremendous technical contributions. He has authored/coauthored over 100 technical papers in journals and conference proceedings, and 20 invention patents. He received 2001 and 2012 National Awards for Science and Technology Progress, China, the 2015 National Award for Technological Invention, China, and the 2014 Distinguished Young Scholar Award of National Natural Science Foundation, China.

Dr. Chen is a fellow of the China Institute of Electronics (CIE), a fellow of the China Institute of Communications (CIC), and a senior member of the IEEE. He is the editor of the *IEEE Network* and the *IEEE Internet of Things Journal*, the guest editor of the *IEEE Wireless Communications Magazine*, and the guest editor of *IEEE Communications Magazine*. He is also the editor of the *China Communications* and the *Journal of Communication*. He was a member of the steering expert group on information technology of the 863 Hi-Tech Research and Development Plan of China from 1999 to 2011.

His current research interests include wireless mobile communication, Internet of Things (IoT) and vehicular network, especially network architecture and mobility management.

Yan Shi received her Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT) in 2007. She is currently a researcher of the State Key Laboratory of Networking and Switching Technology at BUPT. Her current research interests include network architecture evolution, protocol design, and performance optimization of future networks and mobile computing, especially mobility management technology.

Bo Hu is an associate professor at State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications (BUPT), China, since 2006. He received his Ph.D. degree in communication and information system from BUPT, in 2006. His research interests include future wireless mobile communication systems and mobile-driven networking.

Ming Ai received his Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT). He is currently a senior engineer and an R&D project manager of the China Academy of Telecommunication Technology (CATT) and also the leading delegate of CATT for 3GPP TSG SA, TSG CT, and SA2 WG. His research interests include wireless mobile communications and network architecture, mobility management, security, Internet of Things, and standardization activities.

Abbreviations

3GPP	3rd Generation Partnership Project
3GPP2	3rd Generation Partnership Project 2
AAA	Authentication, authorization, and accounting
AC	Access controller
AGW	Access gateway
AH	Authentication header
AHP	Analytic hierarchy process
AKA	Authentication and key agreement
ALT	Automatic link transfer
AMPS	Advanced Mobile Phone System
AMR	Adaptive multi-rate
AN	Access network
ANDSF	Access Network Discovery and Selection Function
AOR	Address-of-record
AP	Access point
API	Application programming interface
AR	Access router
ARQ	Automatic Repeat reQuest
AS	Access stratum
ASCONF	Address configuration change chunk
ASN	Access service network
AU	Application unit
AuC	Authentication center
AuF	Authentication function
B3G	Beyond 3G
BARWAN	Bay Area Research Wireless Access Network
BCCH	Broadcast control channel
BCE	Binding cache entry
BER	Bit error rate
BGCF	Breakout Gateway Control Function

BICC	Bearer-Independent Call Control
BID	Binding Identification Number
BS	Base station
BSC	Base station controller
BSS	Base station subsystem
BSSAP	Base Station System Application Part
BTS	Base transceiver station
BU	Binding update
C2C-CC	Car2Car Communication Consortium
CA	Carrier aggregation
CAC	Call admission control
CAMEL	Customized Applications for Mobile network Enhanced Logic
CC	Call control
CCIR	Consultative Committee on International Radio
CCITT	Consultative Committee of International Telephone and Telegraph
CCO	Cell change order
CDMA	Code division multiple access
CF-SIP	Cross-layer fast handoff for SIP
CH	Correspondent host
CHAP	Challenge handshake authentication protocol
CIP	Cellular IP
CLS	Core logic server
CM	Connection management
CMDP	Constrained Markov decision process
CMIP	Common Management Information Protocol
CN	Core network
CN	Correspondent node
CoA	Care-of address
CoMP	Coordinated multipoint
CoT	Care-of test
CoTI	Care-of test init
CR	Cognitive radio
CRS	Cell-specific reference signal
CS	Circuit switched
CSCF	Call session control function
CSFB	Circuit-switched fallback
CSG	Closed subscriber group
CT2	Cordless Telecommunications 2
CWND	Congestion window
DAD	Duplicate address detection
DAR	Dynamic address reconfiguration
D-CCP	Destination call control point
DCH	Dedicated channel
DDNS	Dynamic domain name system
DECT	Digital European Cordless Telecommunications

DHAAD	Dynamic Home Agent Address Discovery
DHCP	Dynamic Host Configuration Protocol
DHT	Distributed hash table
DL	Downlink
DMM	Distributed mobility management
DMO	Direct mode operation
DNS	Domain name system
DRM	Digital rights management
DRR	Domain root router
DRX	Discontinuous reception
DS	Distribution system
DSMIP	Dual-Stack Mobile IP
DSMIPv6	Dual-Stack Mobile IPv6
DVB-T	Digital Video Broadcasting – Terrestrial
EAP-AKA	Extensible Authentication Protocol for UMTS Authentication and Key Agreement
EAP-SIM	Extensible Authentication Protocol for GSM Subscriber Identity Module
E-DCH	Enhanced DCH
e-ICIC	enhanced inter-cell interference coordination
EID	Endpoint identifier
EIR	Equipment identity register
eMBMS	enhanced MBMS
EMM	Evolved Packet System Mobility Management
eNodeB	Evolved Node B
ENRP	Endpoint Handlespace Redundancy Protocol
EPC	Evolved Packet Core
ePDG	Evolved PDG
EPS	Evolved Packet System
E-RAB	Evolved Radio Access Bearer
ESM	Evolved Packet System Session Management
ESP	Encapsulating Security Payload
ESS	Extended service set
ETSI	European Telecommunications Standards Institute
E-UTRA	Evolved Universal Terrestrial Radio Access
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
FA	Foreign agent
FACH	Forward access channel
FBack	Fast Binding acknowledgement
FBU	Fast Binding Update
FBWA	Fixed broadband wireless access
FCAPS	Fault, configuration, accounting, performance, and security
FCFS	first come, first served
FDD	Frequency-division duplexing
FDMA	Frequency division multiple access

FEC	Forward error correction
FID	Flow identifier
FIFO	First-in first-out
FMC	Fixed mobile convergence
FMIP	Fast MIP
FMIPv6	Fast handover MIPv6
GCoA	Global CoA
GEO	Geosynchronous Earth orbit
GERAN	GSM/EDGE Radio Access Network
GFA	Gateway foreign agent
GGSN	Gateway GPRS support node
GK	Gatekeeper
GMM	GPRS mobility management
GMSC	Gateway MSC
GPRS	General packet radio service
GPS	Global Positioning System
GRA	Grey relational analysis
GSLs	Ground-to-satellite links
GSM	Global System for Mobile Communications
GSS	GPRS Subsystem
GTP	GPRS Tunneling Protocol
GUTI	Global Unique Temporary Identity
HA	Home agent
HAck	Handover Acknowledge
HAWAII	Handover Aware Wireless Access Internet Infrastructure
HBM	Host-based mobility
HCN	Hyper-cellular network
HDB	Home database
H(e)NB	Home eNode B
HetNet	Heterogeneous radio network
HE	Home environment
HF	handoff factor
HHO	Hard handover
HI	Host identity
HI	Handover initiate
HIP	Host Identity Protocol
HiperLAN	High-performance radio local area network
HIT	Host identity tag
HLF	Home location function
HLR	Home location register
HMAC	Hash message authentication code
HMIPv6	Hierarchical Mobile IPv6
HNP	Home network prefix
HoA	Home address
HoL	head-of-the-line

HON	Handover number
HoT	Home test
HoTI	Home test init
HPLMN	Home Public Land Mobile Network
HRPD	High-rate packet data
HSDPA	High-speed downlink packet access
HS-DSCH	High-speed downlink shared channel
HSPA	High-speed packet access
HSS	Home subscriber server
HTP	Hop-To-Paging
HTR	Handover transaction requests
HTTP	Hypertext Transfer Protocol
IBSS	Infrastructure basic service set
ICAO	International Civil Aviation Organization
ICMP	Internet Control Message Protocol
ICN	Information-centric network
ICT	Information and Communications Technologies
ID	Identification
IDC	In-device coexistence
iDEN	Integrated Digital Enhanced Network
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IGMP	Internet Group Management Protocol
IHAP	Inter Home Agent Protocol
IKE	Internet Key Exchange
IKEv2	Internet Key Exchange Version 2
IMEI	International mobile equipment identifier
IMS	IP multimedia subsystem
IMSI	International mobile subscriber identity
IMS-MGW	IMS-Media Gateway
IMT	International Mobile Telecommunications
IP	Internet Protocol
IPCP	Internet Protocol Control Protocol
IPMS	IP mobility management selection
IPsec	Internet Protocol security
IRTF-RRG	Internet Research Task Force Routing Research Group
ISL	Inter-satellite link
ISMP	Inter-system mobility policy
ISR	Idle-mode signaling reduction
ISRP	Inter-system routing policy
IST	Information society technologies
IT	Information technology
I-TCP	Indirect TCP
ITS	Intelligent transport system
ITU	International Telecommunication Union

ITU-D	ITU Development Sector
ITU-R	ITU Radiocommunication Sector
ITU-T	ITU Telecommunication Standardization Sector
IVAN	Intra-Vehicular Area Network
IWF	Interworking function
I-WLAN	Intelligent WLAN
IWS	Interworking solution
IWU	Interworking units
Kbm	binding management Key
LA	Location area
LAI	Location area identifier
LBS	Location-based service
LCA	Least common ancestor
LCoA	on-link CoA
LCoA	Local care-of address
LCS	Location services
LEO	Low Earth orbit
LFN	Local fixed node
LFNs	Local fixed nodes
LGW	Local gateway
LIPA	Local IP access
LISP	Locator/ID separation protocol
LLC	Logical Link Control
LMA	Local Mobility Anchor
LMDS	Local Multipoint Distribution Services
LMN	Local mobile node
LMSD	Legacy mobile station domain
LPN	Low power node
LSS	Location service server
LTE	Long-Term Evolution
LTE-A	Long-Term-Evolution-Advanced
M2M	Machine-to-machine
MAC	Media access control
MAG	Mobile Access Gateway
MAHO	Mobile-assisted handover
MANET	Mobile ad hoc network
MAP	Mobility anchor point
MAP	Mobile Application Part
MAPCON	Multiple-Access Public data network Connection
MBLMS	Movement-based location management scheme
MBMS	Multimedia broadcast/multicast service
MBPS	Measure-based priority scheme
MC	Mobile client
MCDM	Multi-criteria decision making
MCHO	Mobile-controlled handover

MCNA	Mobile controlled network assisted
MDC	Multi-description coding
MDP	Markov decision process
MDT	Minimization of drive tests
ME	Mobile equipment
MEO	Medium Earth orbit
MExE	Mobile Execution Environment
MEXT	Mobility Extensions for IPv6
MF	Mobility function
MFIFO	Modified first-in first-out
MGCF	Media gateway control function
MGW	Media gateway
MH	Mobile host
MIB	Management information base
MIB/SIB	Master information block/system information block
MICS	Media independent command service
MIES	Media independent event service
MIF	Multiple interfaces
MIH	Media independent handover
MIHF	MIH function
MIHU	MIH user
MIIS	Media independent information service
MIMO	Multiple-input multiple-output
MIP	Mobile Internet Protocol
MIP-RR	MIPv4 Region Registration
MIPSHOP	MIPv6 Signaling and Handoff Optimization
MIPv4	Mobile IPv4
MIPv6	Mobile IPv6
MLD	Multicast listener discovery
MM	Mobility management
MMAC	Multimedia mobile access communication
mMAG	moving MAG
MMD	Multimedia Domain
MME	Mobility Management Entity
MMSP	Mobile Multimedia Streaming Protocol
MN	Mobile node
MNG	Mobile Network Gateway
MNN	Mobile network node
MNP	Mobile network prefix
MO	Management Object
MOBIKE	IKEv2 Mobility and Multihoming Protocol
MOCCA	Mobile communication architecture
MONAMI6	Mobile Nodes and Multiple Interfaces in IPv6
MR	Mobile router
MR	Mobile relay

MRFC	Multimedia Resource Function Controller
MRFP	Multimedia Resource Function Processor
MRs	Mobile routers
MS	Mobile station
MS	Mobile server
MSA	Mobility security association
MSC	Mobile Switching Center
mSCTP	mobile Stream Control Transmission Protocol
MSN	Mobile social network
MSR	Mobility support router
MSS	Mobile satellite system
MSS	Mobile satellite service
MT	Mobile terminal
MTC	Machine-type communications
M-TCP	Mobile TCP
MTX	Mobile telephone exchange
M-UDP	Mobile user datagram protocol
MultiMob	Multicast mobility
NAI	Network Access Identifier
NAR	New access router
NAS	Non-access stratum
NAT	Network address translator
NBM	Network-based mobility
NCHO	Network-controlled handover
NCMA	Network controlled mobile assisted
NCoA	New CoA
NEMO	Network mobility
NEMO BSP	Network mobility basic support protocol
NETEXT	Network-based mobility extensions
NETLMM	Network-based localized mobility management
NGI	Next-generation Internet
NGN	Next-generation networks
NIMTC	Network improvements for MTC
N-MAG	New MAG
NMT	Nordic Mobile Telephone
NPoA	Network point of access
NPS	Non-prioritized scheme
NSS	Network subsystem
NTR	New transaction requests
oAR	old access router
OBP	Onboard processing
OBU	Onboard unit
OCEAN	Onboard Communication, Entertainment, and Information
OFDM	Orthogonal frequency-division multiplexing
OLSR	Optimized Link State Routing Protocol

OSA	Open Service Access
OSI	Open Systems Interconnection
P2P	Peer-to-peer
PA	Paging agent
PA	Paging area
PACS	Personal Access Communications System
PAN	Personal area network
PBA	Proxy Binding Acknowledgement
PBU	Proxy Binding Update
PCF	Packet control function
PCO	Protocol Configuration Option
PCoA	Previous CoA
P-CSCF	Proxy Call Session Control Function
PDA	Personal digital assistant
PDCP	Packet Data Convergence Protocol
PDG	Packet Data Gateway
PDLS	Protocol-dependent logic server
PDN	Packet data network
PDN GW	Packet data network gateway
PDP	Packet Data Protocol
PDSN	Packet Data Serving Node
P-GW	PDN gateway
PHS	Public hotspot
PLMN	Public land mobile network
PL-PMIPv6	Packet lossless PMIPv6
p-MAG	previously attached MAG
PMIPv6	Proxy Mobile IPv6
PPP	Point-to-Point Protocol
PR	Preemptive resume
ProSe	Proximity-based services
PrRtAdv	Proxy Router Advertisement
PS	Packet switched
PS domain	Packet-switched domain
PSE	Personal service environment
PS HO	Packet-switched handover
PSK	Phase-shift keying
PSTN	Public switched telephone network
P-TMSI	Packet Temporary Mobile Subscriber Identity
QAM	Quadrature amplitude modulation
QI	Quality indicator
QoE	Quality of experience
QoS	Quality of service
QPS	Queuing priority scheme
R ² CP	Radial reception control protocol
RA	Routing area

RA	registration area
RA	Router advertisement
RACH	Random access channel
RADIUS	Remote Authentication Dial In User Service
RAI	RA identity
RAN	Radio access network
RANAP	Radio Access Network Application Protocol
RAND	Randomly generated number
RAS	Registration, admission, and status
RAT	Radio access technology
RAU	Routing area update
RCoA	Regional CoA
RCP	Reception control protocol
RCS	Reserved channel scheme
RFC	Request For Comments
RIM	RAN Information Management
RLC	Radio link control
RLOCs	Routing locators
RNC	Radio network controller
RPGM	Reference point group mobility
RRC	Radio resource control
RRP	Return routability procedure
RSerPool	Reliable server pooling
RSS	Received signal strength
RSS	Radio signal strength
RSSI	Received signal strength indicator
RSU	Road-side unit
RTO	Retransmission timeout
RTP	Real-time transport protocol
RTT	Radio transmission technology
RTT	Round-trip time
S/MIME	Secure Multipurpose Internet Mail Extensions
SA	Security association
SACK	Selective acknowledgment
SAE	System architecture evolution
SAG	Satellite access gateway
SAMP	Scalable application layer mobility protocol
SAP	Service access point
SATSIX	Satellite-based communications systems within IPv6 networks
SCH	Supplemental Channel
SCTP	Stream Control Transmission Protocol
SD	Safety distance
SDN	Software-defined networking
SDO	Standards Development Organization
SDP	Session description protocol

SG	Study Group
SGSN	Serving GPRS support node
S-GW	Serving Gateway
SHIM6	Site Multihoming by IPv6 Intermediation
SHO	Soft handoff
SIB	System information block
SIGMA	Seamless IP diversity-based generalized mobility architecture
SIM	Subscriber identity module
SIP	Session Initiation Protocol
SIPIHO	SIP Intelligent Handover
SIPTO	Selective IP Traffic Offload
SM	Session management
SMDP	Semi-Markov decision process
SMLC	Serving Mobile Location Center
SMS	Short message service
SMSoSG	SMS over SG
SMVN	Service Mobility Virtual Network
SN	Service network
SDNCP	Sub-network Dependent Convergence Protocol
SNMP	Simple Network Management Protocol
SoLoMo	Social, local, and mobile
SON	Self-optimizing network
SRES	Signed result
SRNS	Serving radio network subsystem
SRP	Service Roaming Protocol
SRP	Selective Repeat Protocol
SRS	Subrating scheme
SRVCC	Single Radio Voice Call Continuity
SS7	Signaling System NO.7
S-SCTP	Secure-SCTP
SSID	Service set ID
SSM	Supplementary service management
STA	Station
SVC	Scalable video coding
TA	Tracking area
TACS	Total Access Communication System
T-ADS	Terminating access domain selection
TAG	Terrestrial access gateway
TAL	Tracking area list
TAU	Tracking area update
TBLMS	Time-based location management scheme
TCP	Transmission Control Protocol
TCP/IP	Transmission Control Protocol/Internet Protocol layered architecture
TCP-R	TCP redirection

TDD	Time-division duplexing
TDMA	Time division multiple access
TD-SCDMA	Time division-synchronous code division multiple access
TEID	Tunnel endpoint ID
TETRA	Terrestrial Trunked Radio
TeleMIP	Telecommunications-Enhanced MIP
TISPAN	Telecommunications and Internet converged Services and Protocols for Advanced Networking
TLMR	Top-level mobile router
TLS	Transport layer security
TMA	TeleMIP Mobile Agent
TMSI	Temporary Mobile Subscriber Identity
TOPSIS	Technique for order of preference by similarity to ideal solution
T-PDU	Tunnel Protocol Data Unit
TV	Television
TWAG	Trusted WLAN (wireless local area network) gateway
TWAN	Trusted WLAN (wireless local area network)
UA	User agent
UAC	User agent client
UAS	User agent server
UDP	User Datagram Protocol
UE	User equipment
UICC	Universal Integrated Circuit Card
UMTS	Universal mobile telecommunications system
UNA	Unsolicited Neighbor Advertisement
URI	Uniform resource identifier
USAT	Universal Subscriber Identity Module Application Toolkit
USIM	Universal Subscriber Identity Module
UTRA	Universal Terrestrial Radio Access
UTRAN	Universal Terrestrial Radio Access Network
VANET	Vehicular ad hoc network
VDB	Visiting database
VHE	Virtual home environment
VIP	Virtual Internet Protocol
VLF	Visitor location function
VLR	Visitor location register
VMN	Visiting mobile node
VoIP	Voice over IP
VPN	Virtual private network
WAG	WLAN Access Gateway
WAN	Wide area network
W-APN	Wireless access point name
WBA	Wireless Broadband Alliance
WCDMA	Wideband CDMA
WEI	Word error indicator

WFA	Wi-Fi Alliance
WG	Working group
WIDE	Widely integrated distributed environment
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless local area network
WLL	Wireless local loop
WM	Wireless medium
WMAN	Wireless metropolitan area network
WPAN	Wireless personal area network
WRIX	Wireless Roaming Intermediary eXchange
WRIX-d	WRIX-Data Clearing
WRIX-f	WRIX-Financial Settlement
WRIX-i	WRIX-Interconnect
WSN	Wireless sensor network
WWAN	Wireless wide area network

Part I
Fundamental Concepts and Principles

Chapter 1

Introduction

This chapter provides an overview of the origination, history, and research background of mobility management technology. Definitions and classifications are introduced with technical terminologies. Standardization activities in 3GPP, IETF, ITU, IEEE, ETSI, WFA, and WBA are summarized briefly, which is followed by motivation and outline of this book.

1.1 Background and Evolution

1.1.1 History

Mobility management, originated from mobile communication systems (i.e., cellular systems), has always been a principal capability of subsequent evolving networks. Cellular communication systems have complete functions of mobility management. More mobility requirements for the Internet emerged with the development of broadband wireless access technologies. Dedicated solutions such as Mobile Internet Protocol (Mobile IP, or MIP) [1, 2] were proposed by the Internet Engineering Task Force (IETF) for mobility management in the Internet. Other protocols with mobility support capabilities were also proposed thereafter. The presence of the mobile Internet signifies the convergence of mobile communication systems with the Internet and calls for increasingly abundant mobility requirements with comprehensive functions of mobility management. Even for future wireless ubiquitous networks, the “5W” communications requirement (i.e., communications by whoever, whenever, wherever, by whatever methods while on the move, with whomever) means that mobility management remains a challenging issue.

Mobile communication is a communication between two mobile objects or between one mobile object and one stationary object [3] (i.e., at least one party of the communication is in motion while the information transmission and exchange

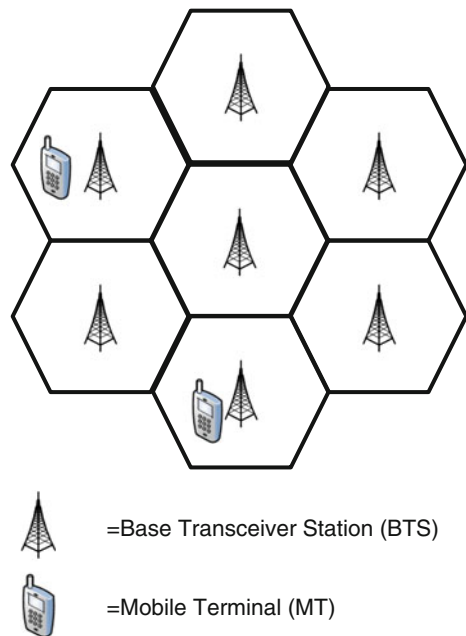
take place). Although the ability to communicate with people on the move has evolved remarkably since Marconi demonstrated the ability of radio to provide continuous contact with ships, the early mobile communication systems were actually simple wireless communications without any mobility management functions until AT&T Bell Labs and other telecommunication companies developed cellular radiotelephony during the 1950s and 1960s. The proposed system aimed to solve the problem of frequency reuse and to improve the spectrum efficiency. The area covered by the mobile communication system was partitioned into adjacent cells, with a base station (BS) in each cell to provide radio access for all users within the cell, as shown in Fig. 1.1.

It is the introduction of cells that led to a series of problems, including location management, handover control, roaming, registration, and authentication of mobile users [4].

Location Management

Location management is a technical requirement introduced by mobile stations in idle mode [4]. In cellular systems, the attaching cell changes when the user moves. The system thus requires a location management function to locate mobile users (i.e., to determine the cell in which the user currently resides). The call setup procedure can therefore be implemented based on the location information. There are three traditional methods that are used for location management: (i) broadcast paging only, which incurs a large paging overhead and large call setup latency; (ii) location update at cell level only, with a large location update overhead which is

Fig. 1.1 Example of cellular network



insupportable in the small cell scenario; and (iii) a compromise solution based on the location area concept. The location area is defined as a set of cells. Each cell belongs only to a single location area. Location updates are needed only when the user moves across different location areas, while paging is only conducted within the location area. This eclectic solution is intended to develop a trade-off between the paging and the location update overheads [4].

Handover Control

Handover control is a technical requirement introduced by mobile stations in the active mode, where the mobile station is in ongoing communication. Handover is required to transfer the ongoing communication from one cell to another to avoid communication interruption for the active mode users. The handover procedure is composed of two stages. The first stage is the handover preparation stage, which is used to detect a change of cells in the active mode. The second stage is when the handover is performed, by transferring the ongoing communication from one channel in the original cell to another channel in the new cell with minimal effects on the users [4].

Roaming, Registration, and Authentication

Roaming, registration, and authentication are requirements for commercial operation and charging. Unlike fixed networks, where the user chooses a permanent stationary network to obtain services, the different networks provide services to mobile terminals according to their locations. Roaming is a general term referring to the extension of the communication service in a location that is different from the home location where the service was originally registered. Roaming requires technical and management support, including charging, registration, and mutual permissions for different mobile terminal types between different telecom operators, location information transfer interfaces, and common access interfaces [4].

The location management, handover control, roaming, registration, and authentication requirements mentioned above are essential issues for mobility management in mobile communication systems. For example, the Global System for Mobile Communications (GSM) standards define the location database, the location area, paging, location updates, and handover procedures. Along with the development of the mobile communication systems themselves, mobility management remains absolutely essential to mobile communication techniques and standards at all times. Table 1.1 summarizes the telecommunication systems that use mobility management technology.

In the Internet field, mobility management has attracted lots of attentions because of the advancement in terminal technology and mobility requirements for broadband wireless access technology. Traditional fixed access technology could not provide anytime and anywhere access to the Internet. Mobile IP (MIP) is the dedicated solution defined by the IETF for mobility management for Internet access. The protocol has been extended from host mobility to network mobility, from macro-mobility to micro-mobility, and also for performance improvement.

Table 1.1 Telecommunication systems that use mobility management technology

Communication system type	Example systems
Cellular mobile communication systems	1G, 2G, 3G, and 4G systems
Satellite mobile communication systems	Low Earth orbit (LEO) systems (e.g., the Iridium system and the Globalstar system) Medium Earth orbit (MEO) systems Geostationary Earth orbit (GEO) systems
Trunked mobile communication systems	Integrated digital-enhanced network (iDEN) Terrestrial trunked radio (TETRA)
Cordless telephone systems	Cordless telecommunications (2CT2) Digital European cordless telecommunications (DECT) Personal access communications system (PACS)

The mobility support capability is also defined at the other layers (e.g., the mobile Stream Control Transmission Protocol (mSCTP) [5] at the transport layer and the Session Initiation Protocol (SIP) [6] at the application layer). The technical principles are also extended. For example, Proxy MIP [7] is a network-based mobility management technology, rather than a host-based solution such as Mobile IP. Host Identity Protocol (HIP) [8] is another example that follows the ID/locator separation principle. Mobility management has actually become one of the most challenging issues in Internet.

When mobile data communication requirements grew rapidly, the users desired access to the Internet from mobile devices such as smartphones, laptops, or tablets. These requirements spurred the beginning of the mobile Internet. Heterogeneity in the access technologies produces new challenges for mobility management. For example, evolution from a horizontal handover to a vertical handover brings asymmetry, active, and dynamic features that must be considered as part of the handover decision and implementation. Some interoperability mechanisms are also necessary to handle the gaps between the different access technologies for mobility-related issues such as quality-of-service (QoS) mapping and service adaptation.

For ubiquitous networks, various access networks will cooperate to provide ambient communication environments. Non-infrastructure networks such as the mobile ad hoc networks (MANETs) and the wireless sensor networks (WSNs) will be integrated for more intelligent and flexible communications. Considering the distributed dynamic topology and ad hoc features of these networks, topology management and routing management should be included in the mobility management. Also, the abundant services and communication modes, such as machine-to-machine (M2M) communication and vehicular networks, present different mobility models and different service traffic models, and thus bring new challenges for mobility management.

Even for future networks following a clean-slate design approach, mobility management will still be an absolutely necessary network capability. The mobility support capability must be embodied in the network architecture design, the identification system design and other network function designs such as routing and resource management.

In summary, with the evolution of network technologies, mobility management has also become one of the key technologies in both the telecommunications field and the Internet field, from traditional Internet access to mobile Internet access, from homogeneous networks to heterogeneous networks, from infrastructure-based networks to non-infrastructure networks, and even in future ubiquitous networks. Indeed, mobility has become a principal feature of network communications. Mobility management is therefore no longer only a specific issue in a specific network, but has become a general technology covering service, network, and terminal aspects.

1.1.2 Research Background

As described earlier, the mobility requirements of the users are derived from “5W” communications. Mobility management is the technology used to achieve these goals and has become one of the most challenging problems for future network and mobile computing. The research background of mobility management, from its origins in cellular communication systems to the Internet field, to mobile Internet and future ubiquitous networks, and from being only a single network technology in a specific network environment to becoming an essential and general feature of networks and communications, has progressed through the evolutionary changes shown in Fig. 1.2.

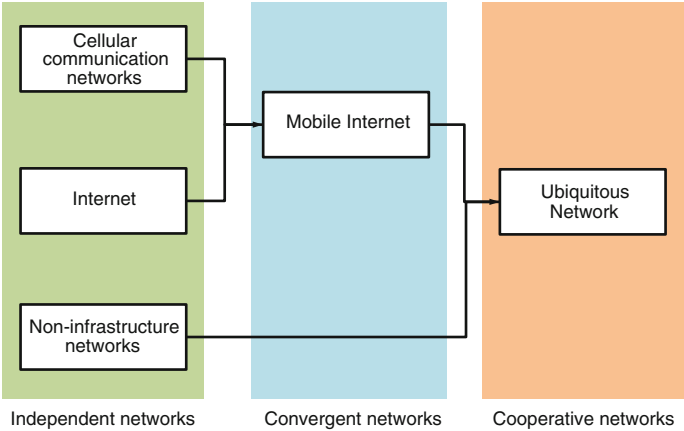


Fig. 1.2 Research background evolution of mobility management

As shown in Fig. 1.2, the relationships between the different networks change from independent to convergent and finally to cooperative. How services were provided based on the network and how the users used these services through the terminals also changed accordingly. For example, for independent networks, specific services are deployed on a certain network and the users can access these services with dedicated terminals. At that point, mobility management was only considered within a particular network. For the ubiquitous network, cooperation between the different networks is introduced. Heterogeneous networks with different features and capabilities cooperate to provide ubiquitous network services for the users and the users can use these personal services with reconfigurable terminals. In this case, mobility management is an important capability that covers service, network, and terminal aspects.

With this evolutionary research background, the requirements, meanings, goals, protocols, and technologies of mobility management have developed along with new features. Mobility management has also become tightly related to wireless access control, radio resource management, and session management technologies.

As discussed above, the history of mobility management covers cellular communication systems, the Internet, and the converged mobile Internet. Figure 1.3 depicts typical communication systems with mobility management entities. Figure 1.3a presents a typical cellular system, where the mobility management-related network entities [e.g., home location register (HLR) and visitor location register (VLR)] are also presented. This is the scenario of the original cellular mobile communication systems. Figure 1.3b illustrates a typical Mobile IPv4 network with mobility management entities such as home agent (HA) and foreign agent (FA). We can easily find some common technical features for mobility management in these different networks, although the networks belong to different fields with different architectures. Summarization and abstraction of the similar entities, functions, and technical principles among the different network architectures would be very useful. Therefore, the protocol reference model and the network reference model of mobility management are introduced in Chap. 2, based on the in-depth research into mobility management of the authors.

1.1.3 Evolution Directions

The coexistence of various heterogeneous access networks, new services, and application requirements, along with numerous different types of mobile terminals, raises new challenges for mobility management technology.

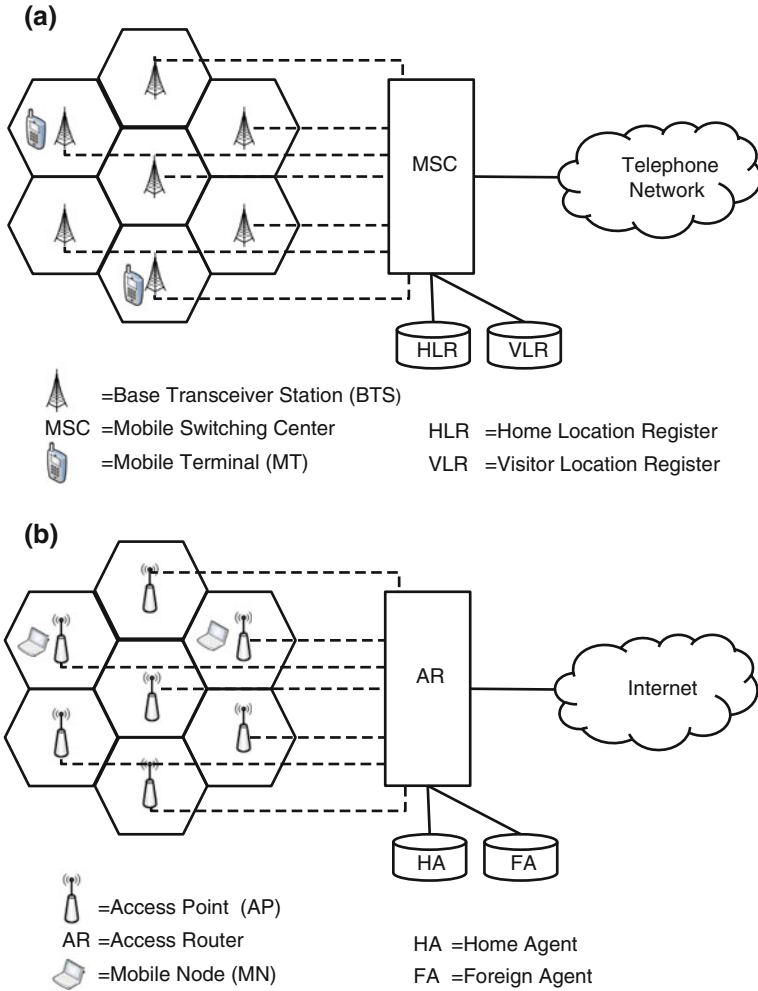


Fig. 1.3 Typical communication system with mobility management entities. **a** Original cellular system, **b** Mobile IPv4 network

The general requirements for future mobility management technology include the following:

(1) Provision of comprehensive mobility support

Traditional mobility management technologies only support terminal mobility. Future mobility management technologies should provide comprehensive mobility support, including support for personal mobility, service mobility, and network mobility in addition to that for terminal mobility. Besides, it should support mobility management for non-infrastructure networks, e.g., support of ad hoc networks [9].

(2) General purpose technology

Future mobility management technology should be independent of the access technologies used, independent of the IP versions used (i.e., supporting both IPv4 and IPv6), independent of the address types used (supporting both public and private addresses), and able to work in harmony with IP-based networks [9].

(3) Provision of the main control functions

The main control functions of future mobility management include security mechanisms, location management [9], handover control, and interoperability control. The security mechanisms are used to provide data privacy, authentication, signal integrity, and mobile service incontestability. The location management function is responsible for tracking and maintaining the location of the user/terminal whenever they move. The handover control function is used to provide session continuity. Interoperability control is the critical function for mobility across heterogeneous networks, to handle gaps in some of the mechanisms of the different networks.

(4) Integration with other technologies

Mobility management should not be considered separately. In many cases, it must be integrated with other relevant technologies. For example, handover control is often jointly designed with resource management for resource optimization. Multi-hop routing and rerouting could also be considered as a necessary part of the handover process in wireless heterogeneous multi-hop networks.

1.2 Definitions

1.2.1 Mobility

Mobility refers to the ability of the user or other mobile entity to communicate and access services, irrespective of any changes in the location or in the technical access environment [9]. Accordingly, mobility management is the set of functions that are used to provide mobility support.

The relationship between “mobility” and “wireless” should be clarified. Actually, there is no inherent binding relationship between “mobility” and “wireless”, although the mobility originates from the wireless communication systems. We use “wireless network” to refer to the various wireless access technologies, where the users access the network via wireless links, but without requisite mobility support. For example, the wireless local loop (WLL) and fixed broadband wireless access (FBWA) technologies are wireless network technologies but they do not have any mobility support functions. In contrast, we use “mobile network” to emphasize the mobility support capability, but the access link can be either wireless or wired. Inevitably, when these two features are integrated, that provides a mobile wireless network.

1.2.2 Classification

According to various criteria, the mobility can be classified into various types, as shown in Table 1.2.

Classification According to Mobility Support Level

Based on the mobility support level (i.e., the service continuity), mobility can be classified into seamless mobility or nomadic mobility. In seamless mobility, the user/terminal can change the network access point without interrupting the current service session. In nomadic mobility, the service session will be stopped completely and will start again when the network access point is changed. The essential difference between seamless mobility and nomadic mobility is therefore whether or not they support service continuity (i.e., whether they support handover or not). A more detailed comparison between seamless mobility and nomadic mobility is shown in Table 1.3.

Table 1.2 Classifications of mobility

Classification criteria		Classifications
Mobility support level		Seamless mobility
		Nomadic mobility
Mobility object		Terminal mobility
		Network mobility
		Personal mobility
		Service mobility
Mobility range	IP field	Macro-mobility
		Micro-mobility
	Telecommunication field	Intra-AN (access network) mobility
		Inter-AN mobility
		Inter-CN (core network) mobility
Access technology types involved		Horizontal mobility
		Vertical mobility
Network features		Mobility in infrastructure-based networks
		Mobility in non-infrastructure networks

Table 1.3 Comparison between seamless mobility and nomadic mobility

	Seamless mobility	Nomadic mobility
Communication while moving	Maintained	Stopped
Current session	Continuous	Interrupted
Handover	Supported/necessary	Not supported/unnecessary
Suitable for real-time service?	Yes	No
Suitable for loss sensitive service?	Yes	No

Another terminology that is similar to nomadism is roaming. Roaming is the ability of users to access services according to their user profiles while moving outside their subscribed home network (i.e., by using an access point to a visited network). Roaming is similar to nomadism, but they are defined from different viewpoints. A comparison between roaming and nomadism is presented in Table 1.4.

Given the different mobility support levels, we can define three wireless access types: fixed access, nomadic access, and mobile access. These types describe the mobility support capabilities of the wireless access networks, including whether access is permitted from different geographical locations and whether session continuity is possible when moving.

In fixed access, it is assumed that the user device is fixed at a single geographical location for the duration of the network service subscription. The user can be connected to or disconnected from the network but without handover support.

In nomadic access, the user device can access the network from different geographical locations. However, for the duration of the service session, the user is required to be at a fixed location because of the absence of handover support. If the user moves to another location, then the ongoing session will be terminated and then re-established.

In mobile access, the user device can both access the network from different geographical locations and maintain ongoing service sessions when moving. Handover is supported to provide session continuity. Comparisons of fixed access, nomadic access, and mobile access are listed in Table 1.5.

Classifications Based on Mobility Objects

Regarding the mobility objects, mobility can be classified into terminal mobility, network mobility, personal mobility, and service mobility.

Terminal mobility (also known as host mobility [10]) is the type of mobility for scenarios where the same terminal equipment is moving or is used at different locations, requiring the terminal to have the ability to access services from different

Table 1.4 Comparison between roaming and nomadism

	Field	Viewpoint	Focus	Solution
Nomadism	Internet field	From a technical viewpoint, emphasizing access point changes caused by mobility	Effects on network access and session	Technical solutions are more important
Roaming	Cellular mobile communication field	From a commercial viewpoint, emphasizing service environment changes caused by mobility	Effects on service access and charging	Operational solutions and agreement between operators are more important

Table 1.5 Comparisons of fixed access, nomadic access, and mobile access

Access type	Locations/mobility	Handover	Examples
Fixed access	Single/stationary	No	Local multipoint distribution services (LMDS) IEEE 802.16d
Nomadic access	Multiple/stationary	No	IEEE 802.16e Wireless local area network (WLAN)
Mobile access	Multiple/moving	Yes	Cellular mobile communication system IEEE 802.16 m

locations and while in motion. The network is required to have the capability to identify and locate that terminal, and to maintain the user’s ongoing communication while the terminal is in motion [9].

Network mobility refers to the ability of the network, where a set of fixed or mobile nodes are networked together, to change, as a group, the point of attachment to the corresponding network upon the movement of the network itself [9, 10].

Personal mobility is the type of mobility for scenarios where the user changes the terminal used for network access at different locations. The ability of the user to access services at any terminal on the basis of a personal identifier and the capability of the network to provide those services are delineated in the user’s service profile [9].

Service mobility is the type of mobility that is applied for a specific service. This involves the ability of a moving object to use the particular (subscribed) service, irrespective of the locations of the user and that of the terminal that is used for that purpose [9]. Service mobility includes service portability, service environment portability, and personal service subscription portability across the available domains, operators, and terminals.

Of the above mobility types, personal mobility and service mobility are often referred to as types of high-level mobility, because they encompass the various facets and targets of user-related mobility. Communication reachability, service personalization, and connectivity are the major factors involved in these mobility types, while communication through the same terminal, through the different network interfaces of a terminal, and through different terminals are the related scenarios.

Classification According to Mobility Range

Regarding the mobility range, the IETF defines macro-mobility and micro-mobility in the IP field [10], while the ITU Telecommunication Standardization Sector (ITU-T) defines intra-AN mobility, inter-AN mobility, and inter-CN mobility in the telecommunications field [9].

Macro-mobility, which is also called global mobility, is the mobility over a large area and usually means the mobility across IP domains [10]. Micro-mobility, which is also called local mobility, is the mobility over a small area and usually refers to the mobility within an IP domain [10].

Intra-AN mobility, inter-AN mobility, and inter-CN mobility refer to the mobility within an AN, across different ANs belonging to the same CN and across different CNs, respectively [9]. These definitions from the ITU-T are based on the network architecture shown in Fig. 1.4. Examples of the different mobility types defined above can also be found in Fig. 1.4. The intra-AN mobility example is marked “3” in Fig. 1.4. 2a and 2b are the examples of inter-AN mobility, where 2a illustrates the mobility between the same types of AN and 2b illustrates the mobility between different types of AN. The mobility between CN1 and CN3 is inter-CN mobility, and is marked as “1” in the figure.

Classification According to the Access Technology Types Involved

Mobility can be classified into horizontal mobility and vertical mobility, according to the access technology types involved. Horizontal mobility refers to the mobility within the same access technology [10]. Vertical mobility refers to the mobility between different types of access technology [10].

Classification According to the Network Features

According to the features of the networks involved, the mobility can be classified into mobility in infrastructure-based networks and mobility in non-infrastructure networks. In the infrastructure-based networks, mobility support requires operations at the mobile terminals and/or the network infrastructure entities. Some mature mobility management technologies have been designed for infrastructure networks (e.g., mobility in cellular mobile communication networks and Mobile IP networks). In non-infrastructure networks, no network infrastructure entities exist to

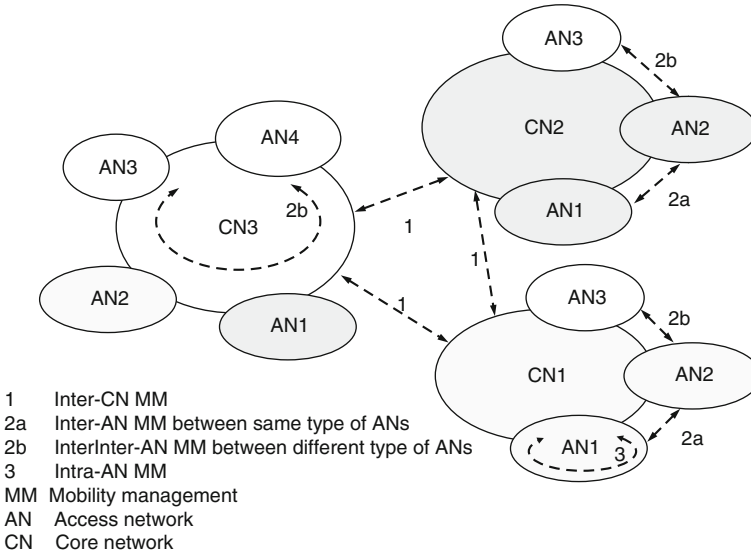


Fig. 1.4 Classification of mobility based on the mobility range in ITU-T [9]

provide mobility support. Mobility management in non-infrastructure networks is often bundled together with topology management, routing, and location management technologies. Mobility in wireless ad hoc networks and WSNs could be taken as examples of this type of management.

A comparison of these two mobility types is shown in Table 1.6.

1.2.3 Related Terminologies

Along with the mobility-related definitions and classifications described above, some other terminologies used in mobility management are also introduced in this section.

Handover

Handover is an important terminology in mobility management. It is the most important technology supporting session continuity, where session interruption and data loss are kept under a certain threshold so that the real-time service remains continuous when the access point changes. As mentioned above, the ability to support handover is the essential difference that separates seamless mobility from nomadic mobility.

For future network environments with ubiquitous, heterogeneous and cooperative features, the handover types will show obvious diversity accordingly. Based on the mobility range, handover could be classified into intranetwork handover and internetwork handover. Horizontal handover and vertical handover are types of handover that depend on whether the access technology types involved are homogeneous or heterogeneous [11, 12]. Based on the handover performance requirements, handover could also be classified into fast handover, smooth handover, and seamless handover [10]. When considering the handover procedure itself (i.e., whether the old link is to be released before or after the new link is established), we can have hard handover, soft handover, and softer handover [13].

Table 1.6 Comparison of mobility between infrastructure-based networks and non-infrastructure networks

Network type	Infrastructure-based	Non-infrastructure
Features	Operations required at the mobile terminals and/or network infrastructure entities	No network infrastructure entities exist to provide mobility support
Mobility management	Mature protocols designed	Not mature, often bundled with topology management, routing and location management technologies
Example networks	Cellular communication system Mobile IP network	Wireless ad hoc network Wireless sensor network

Rescue handover, confinement handover, and traffic handover are handover types that depend on the purpose of the handover [4]. When we consider handover necessity, handover could then be classified into forced handover and unforced handover [14]. Depending on whether user control is permitted or not, we then have active handover and passive handover [15]. More detailed explanations will be provided for all of these types of handover in subsequent chapters.

Another word, handoff, is also often used in the technical literatures. It has the same meaning as handover, and these two words can be used interchangeably. However, these two terms originated from different cellular communication systems. “Handover” came from GSM and its evolving systems and are used in Europe, while “handoff” came from code division multiple access (CDMA) and its evolving systems and are used in the USA. Although the terms are often used interchangeably, some standardization organizations prefer to retain use of their conventional terminologies. For example, 3rd Generation Partnership Project (3GPP) documents based on GSM often use “handover” while 3rd Generation Partnership Project 2 (3GPP2) documents based on CDMA often use “handoff.”

Location Management

Location management is another important terminology and also relates to critical control functions in mobility management. Location management covers tracking, recording, updating, and finding location information for the mobility objects. Another two related terminologies are location update and paging, which are critical control functions in location management. In the location update procedure, the mobility objects report location changes to the network system. Paging is the procedure used by the system to find the current location of each mobility object.

Service Access Point

The service access point is the first network entity to provide network access, number/address adaptation, and related control functions to the users/terminals. The service access point could be different entities in different communication systems, such as the base station subsystem (BSS) in cellular mobile communication systems and the access point (AP) in wireless local area networks (WLANs).

A change in access point may occur for various reasons, such as geographical location changes of the mobility object because of physical movement, changes in the access technologies, or wireless signal strength degradation. A change in the access point may often trigger the location update and handover procedures.

Mobile Node, Mobile Host, Mobile Terminal, Mobile Station, and User Equipment

Mobile node (MN), mobile host (MH), mobile terminal (MT), mobile station (MS), and user equipment (UE) are terms that are often used in the mobility management-related literatures. They can be used interchangeably to refer to mobile terminal devices such as laptops, personal digital assistants (PDAs), smartphones, tablet computers (e.g., the iPad produced by Apple), or PCs. These terminologies originated and are often used in different fields, as shown in Table 1.7.

Table 1.7 Origins of terms MN, MH, MT, MS, and UE

Term	Origin
Mobile station (MS)	Originated from mobile communication systems for mobile transceivers located in automobiles, trucks, boats, and other vehicles Often used in cellular systems to refer to the mobile terminal devices
Mobile host (MH)	Generally used in the Internet field
Mobile node (MN)	Used in the Internet field, especially for Mobile IP and related protocols
Mobile terminal (MT)	Generally used in the mobile communications field
User equipment (UE)	Used in cellular mobile communication systems, and especially used in 3GPP discussion of the LTE (long-term evolution) system

Table 1.8 Comparison between discrete mobility and continuous mobility

	Mobile terminal communication state	Active communication or not?	Principal mobility management technology
Discrete mobility	Idle mode	No	Location management
Continuous mobility	Active mode	Yes	Handover control

Discrete Mobility and Continuous Mobility

Discrete mobility and continuous mobility are defined in ITU-T Q.1703 [16]. Continuous mobility is the ability of a mobile user/terminal/network to change location while the media streams remain active and it requires handover support capability. Discrete mobility refers to the ability of a mobile user/terminal/network to make discrete changes of location (i.e., to change location while no media streams are active). Discrete mobility and continuous mobility are defined for mobile terminals in idle mode and in active mode, respectively, as shown in Table 1.8. In fact, discrete mobility and continuous mobility could also be identified as nomadic mobility and seamless mobility, respectively.

1.3 Standardization Activities

1.3.1 3GPP

The 3GPP unites six telecommunications standards bodies, known as “organizational partners,” and provides their members with a stable environment for producing the highly successful Reports and Specifications that define 3GPP technologies [17].

These technologies are constantly evolving through what have been known as generations of commercial cellular and mobile systems. 3GPP was originally a standards partnership for the evolution of GSM systems toward the third generation. However, since the completion of the first long-term evolution (LTE) and the evolved packet core (EPC) specifications [18], 3GPP has become the focal point for mobile systems development beyond 3G.

From 3GPP Release 10 onward, 3GPP is compliant with the latest ITU-R (ITU Radio communication Sector) requirements for International Mobile Telecommunications (IMT)-Advanced “Systems beyond 3G.” The standard now allows for operation at peak speed of 100 Mbit/s for high mobility communication and 1 Gbit/s for low mobility communication.

3GPP-defined radio access technologies (e.g., GSM, general packet radio service (GPRS), wideband CDMA (WCDMA), time division synchronous CDMA (TD-SCDMA), and LTE) and their corresponding network architectures, mobile terminals, and services (e.g., voice, short message service (SMS), GPRS services, and IP multimedia subsystem (IMS)-related services) are the major focuses of the research performed by the 3GPP community.

Mobility management is one of the most important technologies developed by 3GPP. According to the layered network structure, the mobility management works at three levels: the radio access network level, the core network level, and the service level (e.g., IMS).

From a radio access network perspective, idle mode mobility and connected mode mobility are two important issues faced by every generation of radio access technology (RAT), from GSM to LTE. The terminal is controlled by the network to perform wireless signal measurements, cell selection and re-selection, and handover from one cell to another cell. Initially, there was only intra-RAT mobility because the terminals only support one RAT, which is typically GSM. However, with the emerging new RATs such as universal terrestrial radio access (UTRA) and evolved UTRA (E-UTRA) technologies, inter-RAT mobility must be supported.

Within the 3GPP-defined network (i.e., public land mobile network (PLMN)) for 3GPP RAT, mobility management for an idle mode terminal consists of location management and reachability management. The former lets the network know approximately where the terminal is located, even if the terminal is moving around, and the latter enables the network to find the terminal precisely when terminating services such as voice calls and SMS need to reach the user.

For mobility management of a connected mode terminal, handover is the most important issue, to enable service continuity when the terminal moves from one cell to another cell of the same RAT (intra-RAT) and to change its attachment points from one RAT to another (inter-RAT) and from one PLMN to another (inter-PLMN). The above is the cornerstone for most 3GPP services, ranging from basic services, such as voice calls, SMS, and GPRS, to advanced services such as IMS.

Interworking with other types of radio access technology, (which are normally called non-3GPP RATs) such as Wi-fi, WiMax, and CDMA 1x, is one of the most important challenges that 3GPP has encountered to date. Regarding the mobility

management technology used for terminals that are equipped with 3GPP RAT and non-3GPP RAT, 3GPP has already defined the following functions [19]: access network discovery and selection function (ANDSF), IP mobility management selection (IPMS), functions which fulfill seamless handover between 3GPP RAT and non-3GPP RAT, and functions which fulfill non-seamless off-loading from 3GPP RAT to non-3GPP RAT. In its latest work, the 3GPP is to specify solutions for IP flow mobility between 3GPP and non-3GPP access.

Another important challenge for 3GPP is how to enhance the current system, which was designed for human-to-human communications, to accommodate machine-type communications (MTC), also known as M2M communications, and the mobile data communications of smartphones at the same time. These enhancements, which will make the system more efficient, include the efficient transfer of small amounts of data, reduction of terminal power consumption, enabling group-based communications, and reducing the cost of the MTC terminal. Within those enhancements, mobility management is still one of the most important aspects and requires further consideration [20].

3GPP also considers mobile relay (MR) deployment in 3GPP systems [21]. High-speed public transportation is being deployed at an increasing rate. Provision of multiple good-quality services to users on high-speed vehicles is more complex than in typical mobile wireless environments. The research on mobile relays in 3GPP is focused on the high-speed train scenario [21], with the aim of identifying the potential scenarios and requirements, identifying the key properties of the mobile relays and assessing the benefits of mobile relays over those of existing solutions in fast-moving environments. The mobile relay is also considered as an optimization of group mobility in the high-speed train scenario to reduce mobility management overheads and latency. Therefore, mobility-supported architecture extensions and mobility procedures based on MR deployment are also discussed in 3GPP.

1.3.2 IETF

The IETF [22] is an open international organization in the field of Internet and is composed of network operators, service providers, researchers, and network engineers. The IETF is committed to the development and smooth operation of the Internet architecture. Most of the actual work in the IETF is accomplished by its working groups (WGs), and the WGs are divided into several areas according to their different research topics, including routing, transmission, and network security. The use of open standards is the unique feature of IETF when compared with other standardization organizations in the telecommunications field and has spurred the rapid development of the Internet.

Regarding mobility definitions, the SEAMOB (the SEAmless IP MOBility) WG in the IETF defined various IP mobility-related terms in [10]. These terms define different aspects of, and solutions to, support mobility, by comparison with

aspects such as range, overlapping area, performance characteristics, diversity characteristics, state migration, mobility types, and the control mode of the hand-over technology. More definitions and classifications of high-level mobility are discussed in the SIPPING WG [23]. The IETF is always at the forefront of the research on mobility management protocols and solutions.

On mobility management protocols, the IETF has developed some classical protocols at the network layer, the transport layer, and the application layer. At the network layer, the MIP4 WG mainly researches the MIPv4 protocol to support node mobility. The MIP6 WG aimed to enhance the mobility support capability of MIPv6 and solve problems related to creation and testing of the MIPv6 protocol. The NEMO WG was concerned with management of the mobility of an entire network and subnet; it also proposed and maintains a network mobility basic support protocol, the NEMOv6. The NETEXT (network-based mobility extensions) WG proposed PMIPv6 (Proxy MIPv6) as an extension of MIPv6. PMIPv6 follows the network-based mobility management principle, whereas Mobile IP is a host-based mobility management protocol. Mobile Stream Control Transmission Protocol (mSCTP) [5] was proposed at the transport layer for end-to-end mobility with multi-homing support. The SIP [6] is an application-layer protocol for both terminal mobility and high-level mobility support.

The IETF also conducts research on the ID-locator separation technologies with inherent mobility support capabilities. The HIP [8] is a protocol defined by the HIP WG. The HIP defines the host ID to enable separation of the identifier and locator roles of IP addresses and enables continuity of communications when the mobile node's IP address changes. The Locator/Identifier Separation Protocol (LISP) [24] defined by the LISP WG enables separation of the IP addresses into endpoint identifiers (EIDs) and routing locators (RLOCs) from the network side and can realize efficient routing locator aggregation and provide persistent identifiers.

Given the trend for flatter architectures in network evolution, the IETF conducted research into distributed mobility management in the distributed mobility management (DMM) WG. In traditional mobility management solutions, mobility management with centralized mobility anchoring in the existing hierarchical mobile networks is quite prone to non-optimal routing, lack of scalability, single point of failure and attack, and inevitably introduce longer delays and higher signaling loads for the network operations related to mobility management [25]. Thus, the aim of the DMM WG was to match the mobility deployment with current network evolution trends and to overcome the above problems of centralized mobility management solutions. The DMM WG has now documented the requirements, the current practices and gap analysis, the distributed mobility management framework, and the possible mobile network scenarios for DMM, along with some applicable schemes for mobility management protocols.

The MultiMob (multicast mobility) WG provides guidance for supporting multicast in the mobile environment. The group works on extensions of proxy mobile IPv6 (PMIPv6) to improve its capability to handle multicast efficiently [26]. To date, the working group has already documented how the existing mechanisms can be used to support multicast, without any additions or changes to the message

types and parameters specified in PMIPv6. The group has published three RFCs that address deployment for multicast listener support in PMIPv6 [27], Multicast mobility routing optimizations for proxy mobile IPv6 [28] and tuning of the behavior of the Internet Group Management Protocol (IGMP) and multicast listener discovery (MLD) for routers in mobile and wireless networks [29]. Some drafts for handover optimization and sender support in PMIPv6 are also active [30–32].

1.3.3 ITU

The International Telecommunications Union (ITU) is an international organization within the United Nations. The ITU enables government and nongovernment bodies to coordinate telecommunication networks and services globally and comprises three sectors: the standardization sector (ITU-T), the radio communication sector (ITU-R), and the development sector (ITU-D). The ITU has 191 member states to date, including more than 700 department members and department associate members, such as telecommunications operators, equipment manufacturers, financial groups, telecommunications research and development institutions, and international and regional telecommunications groups.

ITU-T is the telecommunications standardization sector of the ITU and was previously known as the Consultative Committee of International Telephone and Telegraph (CCITT). ITU-T makes research and development of all standards in the telecommunications field, except for the radio aspects. At present, the ITU-T has 10 study groups engaged in operational aspects, economic and policy issues, environmental and climate change, broadband cable networks and TV, protocols and test specifications, performance, QoS and quality of experience (QoE), future networks, transport and access, and multimedia and security.

At present, study group 13 (SG13) in ITU-T involves research work into mobility management. ITU-T SG13 is mainly responsible for research on next-generation networking (NGN) requirements, architecture, evolution, and convergence, where mobility support has been identified as one feature of NGN. SG13 also cooperates closely with other related ITU-T SGs and other international standards organizations such as IETF and 3GPP.

ITU-T has currently published a series of standard recommendations on mobility management, as shown in Table 1.9.

The study groups related to mobility management in ITU-T include the following:

- Q9/SG13, which is responsible for research in mobility management and fixed mobile convergence, based on the mobility management requirements established in Q.1706/Y.2801 and Q.1707/Y.2801 and the fixed mobile convergence (FMC) requirements established in Q.1762/Y.2802. The tasks include support of global roaming and seamless mobility, delivery of services within or across networks for both IMT and NGN, and provision of access independent services and energy savings.

Table 1.9 Standard recommendations of ITU-T on mobility management

No.	Title	Description
Q.1706/Y.2801 [9]	Rec. MMR, or mobility management requirement for NGN	Defining the mobility requirements of NGN, including different mobility types in NGN and related functional groups
Q.1707/Y.2804 [33]	Q. MMF, or generic framework of mobility management for NGN	Defining the general mobility management framework, and determining the basic principles for research into two key mobility management functions: location management and handover control
Q.1708/Y.2805 [34]	Q. LMF, or location management framework for NGN	Defining the frameworks for location management and handover management, respectively
Q.1709/Y.2806 [35]	Q. HCF, or handover control framework for NGN	

- Q10/SG13, which covers the mobility management of multi-connections using multiple access technologies, where the main issues include the multi-connection architecture, access network discovery and selection, identification (ID), multimedia telephony, multi-path transmission control, and flow-based service continuity.

ITU-R is the radio communications sector of ITU. Its predecessor was the Consultative Committee on International Radio (CCIR). ITU-R is responsible for research in wireless communication technology and operations and currently has six study groups, covering spectrum management, radio wave propagation, satellite services, terrestrial services, broadcasting services, and science services. Currently, ITU-R focuses on the research into global 4G mobile wireless broadband technology (IMT-Advance) with emphasis on spectrum division methods and special mobility requirements.

1.3.4 IEEE

The Institute of Electrical and Electronics Engineers (IEEE) is a nonprofit professional association headquartered in New York, USA, that is dedicated to advancing technological innovation and excellence for the benefit of humanity. The IEEE and its members inspire a global community through the IEEE’s highly cited publications, conferences, technology standards, and professional and educational activities. IEEE standards affect a wide range of industries including power and energy, biomedical and health care, information technology (IT), telecommunications, transportation, nanotechnology, information assurance, and many more.

The IEEE defined the 802.1-802.20 series standards, in which the wireless access technology standards include 802.11x, 802.15x, 802.16x, and 802.20x. These wireless access technologies with different frequency ranges, data rates, covering radii, and applications are often treated as parts of heterogeneous, convergent, and ubiquitous networks. IEEE 802.22 working group defined a novel wireless air interface (i.e., MAC and PHY) standard based on cognitive radios (CRs) [36]. It is a standard for wireless regional area network (WRAN) using white spaces in the television (TV) frequency spectrum [37].

The IEEE 802.21 working group [38] also defined the media independent handover (MIH) standards. Existing 802 standards do not allow for the detection and selection of network access points in a way that is independent of the access technology type. The increasing number of interfaces on mobile devices makes it difficult to select the suitable access network. The IEEE 802.21 WG is developing standards to enable handover and interoperability between heterogeneous network types, including both 802 and non-802 networks. The WG started in 2004 and has published three standards. It tries to define media independent handover mechanisms for handover optimization. MIH is primarily intended for vertical handovers and helps with handover initiation, network selection, and interface activation. MIH works between the link layer and the network layer. It provides the MIH event service, the MIH command service, and the MIH information service.

1.3.5 ETSI

The European Telecommunications Standards Institute (ETSI) [39] is a nonprofit organization with more than 700 ETSI member organizations coming from 62 countries across 5 continents worldwide. The ETSI produces globally applicable standards for information and communications technologies (ICT), including fixed, mobile, radio, converged, broadcast, and Internet technologies. The ETSI cooperates with other international standard organizations including the 3GPP, the International Organization for Standardization (ISO), and the ITU.

The Telecommunications and Internet converged Services and Protocols for Advanced Networks (TISPAN) group in ETSI has delivered the technical standard TISPAN (Telecommunications and Internet converged Services and Protocols for Advanced Networking) NGN Release 1 [40]: Release Definition, which gives a general introduction to NGN and defines its functional requirements, including some mobility-related concepts. This technical report quotes definitions of various terms including mobility, session continuity, nomadism, roaming, and seamless handover from Q.1706/Y.2801.

Among the related technical reports in TISPAN, NGN Terminology defines NGN concepts and terminology; NGN Functional Architecture Release 1 describes the NGN functional framework; and in Interface Protocol Definition, the roaming interface in NGN is defined.

1.3.6 WFA

The Wi-fi alliance (WFA) is a global nonprofit organization that started with just 6 member companies in 1999, but there are now more than 500 member companies from dozens of countries. The initial goal of WFA was to drive the adoption of high-speed wireless local area networking, which has been achieved already through a combination of technology development, market building, and regulatory programs (e.g., the Wi-Fi CERTIFIED™ program, which was launched in 2000 and helps to ensure that Wi-fi-enabled products deliver the best user experience) [41].

The current focus of the WFA was seamless connectivity (i.e., via the latest and still ongoing Wi-Fi CERTIFIED Passpoint™ program). The aim of this program was to enable seamless connectivity without any user intervention during network discovery and selection, and network access authentication procedures. The security is also enhanced for authentication and connectivity, with a level comparable to that of cellular networks.

1.3.7 WBA

The Wireless Broadband Alliance (WBA) was founded in 2003. Its members are from major fixed telecom operators, top-rated mobile operators (by revenue), and leading technology companies [42].

The aim of the WBA was to ensure an outstanding user experience through the global deployment of next-generation Wi-fi, by technology innovation, interoperability, and robust security improvement, delivering seamless connectivity and optimized network investments.

The WBA has already published some specifications. The two described below are considered very important for global Wi-fi roaming.

- The Wi-Fi Roaming Guidelines [43], which examine the commercial and technical aspects of how to enable Wi-fi roaming, such as the automatic provisioning of connectivity to end users across different service providers' Wi-fi networks.
- The Wireless Roaming Intermediary eXchange [44], which offers a modularized set of standard service specifications to enable commercial roaming between operators. It includes WRIX-i (Interconnect), WRIX-d (Data Clearing), and WRIX-f (Financial Settlement).

1.4 Motivation and Outline

In general, published books in this field either focus on a particular cellular communication system (e.g., GSM, UMTS, WCDMA, or LTE) where mobility management is included as a key technology, or are dedicated to Mobile IP technology.

As described above, mobility management is indeed an important technology when aiming to provide service continuity support. It has evolved to become an essential feature of network communications. Traditional mobility management technologies were designed for a single network or a particular access technology. They could not satisfy the mobility requirements in the future heterogeneous and convergent network environments. Therefore, this book treats mobility management as an individual technology and provides a system study of the technology.

The content is organized according to the protocol reference model proposed by the authors, including a multi-dimensional study of mobility management requirements and technologies, in-depth analysis of some typical technologies based on the protocol reference model, and potential research directions.

The content of this book is arranged as follows:

This chapter has given a system overview of the background, evolution, definitions, classifications, and standard activities in the field of mobility management.

Chapter 2 describes the mobility management reference model proposed by the authors, including a protocol reference model and a network reference model.

Chapter 3 surveys the existing technologies that support different mobility types, such as terminal mobility, personal mobility, service mobility, and network mobility.

Chapter 4 summarizes the fundamental theories applied in mobility management research, including probability theory, stochastic processes, queuing theory, and fuzzy control theory.

Chapters 5–8 analyze the classical mobility management technologies at the data link layer, the network layer, the transport layer, and the application layer, based on the reference model given in Chap. 2 and focus on the technical principles rather than simply give an introduction to the protocol details.

Chapter 9 introduces the future challenges and potential research directions for mobility management technology.

References

1. IETF RFC3344 (2002) IP mobility support for IPv4
2. IETF RFC3775 (2004) Mobility support in IPv6
3. Black Uyless (1998) Emerging communications technology, 2nd edn. Tsinghua University Press, Prentice Hall International, Beijing
4. Mouly M, Pautet MB, Foreword By-Haug T (1992) The GSM system for mobile communications. Telecom Publishing, Washington, DC
5. IETF draft-sjkoh-msctp-01 (2005) Mobile SCTP (mSCTP) for IP Handover Support
6. IETF RFC3261 (2002) SIP: session initiation protocol
7. IETF RFC6279 (2011) Proxy mobile IPv6 (PMIPv6) localized routing problem statement
8. IETF RFC5201 (2008) Host identify protocol
9. ITU-T Q.1706/Y.2801 (2006) Mobility management requirements for NGN
10. IETF RFC3753 (2004) Mobility related terminology

11. Chakravorty R, Vidales P, Patanapongpibul L et al (2003) On inter-network handover performance using mobile ipv6. University of Cambridge Computer Laboratory–Technical Report. <http://www.cl.cam.ac.uk/coms/publications.htm>. Accessed 23 Mar 2014
12. Siddiqui F, Zeadally S (2006) Mobility management across hybrid wireless networks: trends and challenges. *Comput Commun* 29(9):1363–1385
13. UMTS (Universal Mobile Telecommunications System) handover <http://www.umtsworld.com/technology/handover.htm>. Accessed 23 Mar 2014
14. Xie J, Wang X (2008) A survey of mobility management in hybrid wireless mesh networks. *IEEE Netw* 22(6):34–40. doi:10.1109/MNET.2008.4694172
15. Chen Y, Kowalik K, Davis M (2009) MeshScan: performance of passive handoff and active handoff. *IEEE Int Conf Wireless Commun Signal Process WCSP 2009*:1–5. doi:10.1109/WCSP.2009.5371523
16. ITU-T Q.1703 (2004) Service and network capabilities framework of network aspects for systems beyond IMT-2000
17. The 3rd Generation Partnership Project (3GPP) (1999) <http://www.3gpp.org>. Accessed 23 Mar 2014
18. GPP TS23.401 (2007) General packet radio service (GPRS) enhancements for evolved universal terrestrial Radio access network (E-UTRAN) access
19. GPP TS23.402 (2010) Architecture enhancements for non-3GPP accesses
20. GPP TR23.887 (2013) Study on machine-type communications (MTC) and other mobile data applications communications enhancements
21. GPP TR36.836 (2013) Mobile relay for E-UTRA
22. The Internet Engineering Task Force (IETF) (1996) <http://www.ietf.org>. Accessed 23 Mar 2014
23. IETF draft-itsumo-sip-mobility-req-02 (2000) Mobility management in a SIP environment requirements, functions and issues
24. IETF RFC 6830 (2003) The locator/ID separation protocol (LISP)
25. IETF draft-ietf-dmm-requirements-15 (2014) Requirements for distributed mobility management
26. Multicast Mobility (multimob) working group of IETF (2010) <http://datatracker.ietf.org/wg/multimob/charter/>. Accessed 23 Mar 2014
27. RFC 6224 (2011) Base deployment for multicast listener support in proxy mobile IPv6 (PMIPv6) domains
28. RFC 7028 (2013) Multicast mobility routing optimizations for proxy mobile IPv6
29. RFC 6636 (2012) Tuning the behavior of the internet group management protocol (IGMP) and multicast listener discovery (MLD) for routers in mobile and wireless networks
30. IETF draft-ietf-multimob-fmipv6-pfmipv6-multicast-05 (2014) Multicast listener extensions for MIPv6 and PMIPv6 fast handovers
31. IETF draft-ietf-multimob-handover-optimization-07 (2014) PMIPv6 multicast handover optimization by the subscription information acquisition through the LMA (SIAL)
32. IETF draft-ietf-multimob-pmipv6-source-08 (2014) Mobile multicast sender support in proxy mobile IPv6 (PMIPv6) domains
33. ITU-T Recommendation Q.1707/Y.2804 (2008) Generic framework of mobility management for next generation networks
34. ITU-T Recommendation Q.1708/Y.2805 (2008) Framework of location management for NGN
35. ITU-T Recommendation Q.1709/Y.2806 (2008) Framework of handover control for NGN
36. Cordeiro C, Challapali K, Birru D et al (2005) IEEE 802.22: the first worldwide wireless standard based on cognitive radios. In: 2005 First IEEE international symposium on new frontiers in dynamic spectrum access networks, DySPAN, pp 328–337. doi:10.1109/DYSPAN.2005.1542649
37. IEEE 802 LAN/MAN Standards Committee 802.22 WG on WRANs (Wireless Regional Area Networks). <http://standards.ieee.org/develop/wg/WG802.22.html>. Accessed 24 Mar 2014
38. IEEE 802.21, <http://www.ieee802.org/21/>. Accessed 23 Mar 2014

39. European Telecommunications Standards Institute (2008) <http://www.etsi.org>. Accessed 23 Mar 2014
40. ETSI TISpan Specification (2006) NGN Release 1, <http://www.etsi.org/tispan/>. Accessed 24 Mar 2014
41. Wi-Fi Alliance, <http://www.wi-fi.org>. Accessed 24 Mar 2014
42. Wireless Broadband Alliance, <http://www.wballiance.com>. Accessed 24 Mar 2014
43. Wi-Fi Roaming Guidelines, <http://www.wballiance.com/resource-center/specifications/>. Accessed 24 Mar 2014
44. Wireless Roaming Intermediary eXchange (WRIX—i, l, d & f), <http://www.wballiance.com/resource-center/specifications/>. Accessed 24 Mar 2014

Chapter 2

Mobility Management Reference Models

This chapter is dedicated to the mobility management protocol reference model and the network reference model. These models were proposed by the authors based on a summary of various mobility management technologies.

2.1 Protocol Reference Model

The protocol reference model for mobility management is illustrated in Fig. 2.1. The model is composed of three planes: the data plane, the control plane and the management plane. The data plane illustrates the protocol layers that are involved in mobility management, including the physical layer, the data link layer, the network layer, the transport layer and the application layer. The control plane presents the key control functions in mobility management, including the security mechanism, location management, handover control, and interoperability control. The management plane handles the network management-related functions, including configuration management, fault management, performance management, accounting management, and security management. The functions and typical technologies of these planes will be introduced in Sects. 2.1.1, 2.1.2 and 2.1.3.

2.1.1 Data Plane

The data plane in the protocol reference model follows the revised Transmission Control Protocol/Internet Protocol-layered architecture (TCP/IP), which is composed of the physical layer, the data link layer, the network layer, the transport layer, and the application layer, from bottom to top. The physical layer can provide mobility-related physical signal measurements to be used for mobility management

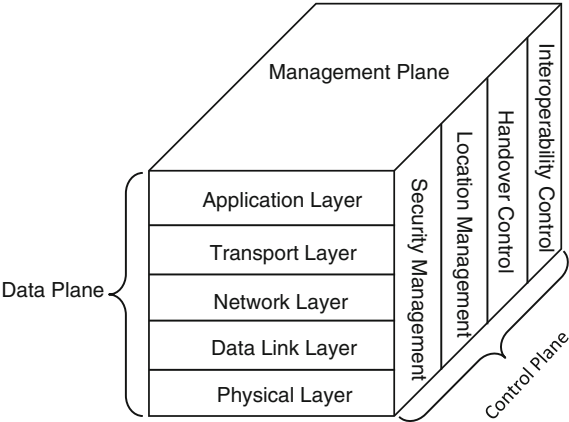


Fig. 2.1 Protocol reference model for mobility management. Reprinted from ref. [1], Copyright 2007, with permission from Journal of communications

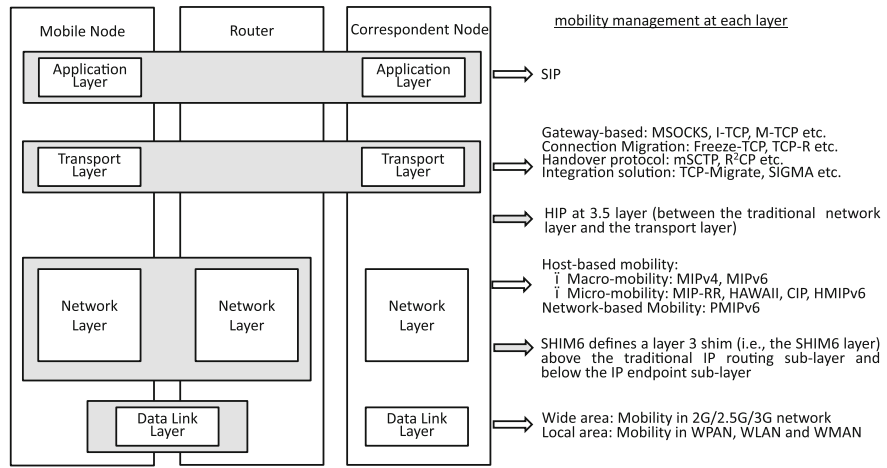


Fig. 2.2 Typical mobility management at the different layers

optimization. Typical mobility management protocols exist at each of the other layers.

Figure 2.2 shows the mobility management protocols and technologies at each layer.

Typical mobility management solutions used at each layer are shown on the right side of Fig. 2.2 and are described for each layer below.

- Mobility management at the data link layer handles the mobility within a subnet area, for example, the mobility in wide area 2G/2.5G/3G cellular mobile communication networks, or mobile access within networks such as the wireless

personal area network (WPAN), the wireless local area network (WLAN) and the wireless metropolitan area network (WMAN).

- Mobility management protocols at the network layer can be classified as host-based mobility supporting or network-based mobility supporting solutions, depending on whether the mobile node participates in mobility management or not. Host-based mobility supporting protocols can be further divided into macro-mobility supporting protocols and micro-mobility supporting protocols.
- The transport layer provides end-to-end mobility support. Mobility management technologies at the transport layer can be classified into gateway-based solutions, connection migration solutions, handover protocols, and integrated solutions.
- The application layer also provides end-to-end mobility support. The Session Initiation Protocol (SIP) is the typical mobility support protocol at the application layer.

These typical protocols and technologies at the different layers will be briefly introduced in Sect. 2.4 and described in more detail in Chaps. 5–8.

Along with the more classical and mature mobility management technologies mentioned above, some new technologies are also proposed based on new technical principles, e.g., ID (identifier)/locator separation solutions with inherent mobility support capability. In general, a permanent identifier associated with dynamically changed locators and the related mapping functions are defined in these solutions. Host Identity Protocol (HIP) [2] and Site Multi-homing by IPv6 Intermediation (SHIM6) [3] are the representative solutions here. HIP introduces a new namespace between the traditional network layer and the transport layer. It defines the host ID to separate the identifier and locator roles of IP addresses. This enables communication continuity when the MN's IP address changes. SHIM6 defines a Layer 3 shim (i.e., the SHIM6 layer) above the traditional IP routing sublayer and below the IP endpoint sublayer. It uses a fixed endpoint identifier above the SHIM6 layer and uses the dynamic IP addresses as locators below the SHIM6 layer. SHIM6 does not define the new namespace in the same way as HIP. Instead, it uses the IP address used in the establishment of the initial session as the fixed endpoint identifier.

For the different mobility types mentioned in Sect. 1.2.2, mobility management technologies at the data link layer and the network layer are often used to support the terminal mobility and network mobility, while the technologies at the transport layer and the application layer often provide personal mobility and service mobility support. Accordingly, terminal mobility and network mobility are called low-level mobility, while personal mobility and service mobility are known as high-level mobility.

Table 2.1 describes the basic functions of each protocol layer in mobility management technologies.

As illustrated in Table 2.1, all of the layers, with the exception of the physical layer, have typical mobility support protocols. Table 2.2 presents a comparison between the typical technologies at the different layers.

Table 2.1 Basic functions of data plane

Protocol layer	Basic functions in mobility management
Physical layer	• Provides mobility management-related physical signal detection and measurement, which can be used for function and performance optimization
Data link layer	• Provides terminal mobility within an IP subnet • Provides necessary information about link status and L2 (Layer 2) handover starting/finishing event notification, which can be used for function and performance optimization
Network layer	• Provides mobility independent of the lower-layer protocols and physical transmission media, and transparent to the upper layers • Mainly supports terminal mobility and network mobility • Provides L3 (Layer 3) handover starting/finishing event notification to the upper layers for handover performance optimization
Transport layer	• Provides end-to-end mobility support
Application layer	• Provides various types of mobility support, especially for high-level mobility (personal mobility and service mobility)

Table 2.2 Comparison of typical mobility protocols at each layer

Protocol layer	Typical technology	Advantages	Disadvantages
Data link layer	Mobility management technology in cellular communication networks	• Supports mobility within an IP subnet, without modifications to network layer, transport layer or application layer	• Mobility is limited to the area of the IP subnet only, i.e., the user could not move from one IP subnet to another without other mobility management technologies
Network layer	MIP	• Provides mobility independent of the lower-layer protocols and physical transmission media and is transparent to the upper layers	• High latency incurred by registration procedure if MN is far from a Home Agent (HA); • Scalability problem, i.e., high signaling overhead incurred by frequent registration when number of mobile nodes (MNs) increases; • Requires deployment of HAs in home network, as well as foreign agent (FA) in visiting network for MIPv4; • Requires modifications to network layer of MN;

(continued)

Table 2.2 (continued)

Protocol layer	Typical technology	Advantages	Disadvantages
			• Unfit for scenario of continuous movement of MN because of performance degradation
Transport layer	mSCTP	• Follows the end-to-end semantics and does not rely on the support of network infrastructures like routers for mobility support	• Requires some modification to the transport layer of the MN; • Cannot provide mobility support to the applications based on other transport layer protocols
Application layer	SIP	• No modifications to transport layer and network layer of MN and correspondent node required; • supports personal mobility and service mobility based on some extensions	• Requires deployment of application-related servers; • Cannot provide session continuity for other application types; • High handover latency

As shown in Table 2.2, the mobility management technologies at each layer have their own particular advantages and disadvantages. Indeed, no individual technology could satisfy the functional and performance requirements of general mobility. Advanced mobility management technology based on cross-layer principles can integrate the advantages of the different layer technologies and thus provide comprehensive mobility support with good performance. This is the reason why the physical layer is included in the data plane, despite the fact that there are no mobility protocols at this layer. Such cross-layer optimization is also illustrated in Table 2.1. The physical layer, the data link layer and the network layer can all provide mobility-related information and event notifications for mobility management performance, and for handover performance optimization in particular.

2.1.2 Control Plane

The basic functions of the control plane in the mobility management protocol reference model include security mechanism, location management, handover control, and interoperability control.

- Security mechanism handles functions related to Authentication, Authorization, and Accounting (AAA) , as well as the user data and privacy protection in mobility.
- Location management is responsible for storing, updating, and retrieving the location information of the mobile objects.
- Handover control enables session continuity when the access point changes.
- Interoperability control is the particular function in mobility management for a heterogeneous network environment driven by the diversity of the access technologies.

For a more detailed description of these critical control functions in mobility management, please refer to Sect. 2.3.

2.1.3 Management Plane

The management plane in the mobility management protocol reference model is in charge of the essential network management functions. This involves the network management protocols and functions. The traditional network management functions, known as fault, configuration, accounting, performance, and security management (FCAPS), are also included here.

(1) Basic management plane functions

Fault management: the process of monitoring and management of abnormal operations, including maintaining the fault log, locating the problems, isolating the problems, diagnostic testing, and fixing the problems, if possible.

Configuration management: the process of finding and setting up critical configurations on the network devices, collecting, storing, modifying and monitoring the configuration information, and providing configuration information to other related systems. This involves initialization and deletion of managed objects, setting up of suitable configurations for regular operations, and collection of status information.

Accounting management: this function involves tracking each individual's usage and grouping of the network resources to ensure that the users have sufficient resources.

Performance management: this function is useful for system/resource performance evaluation, including: collection and monitoring of performance statistics, maintaining the historic records of the system performance, and measurement of the performance of the network hardware, software, and media.

Security management: this refers to the various protection functions for system security, including maintaining the security log, providing audit trails, and sounding alarms, and distribution of security information to other systems.

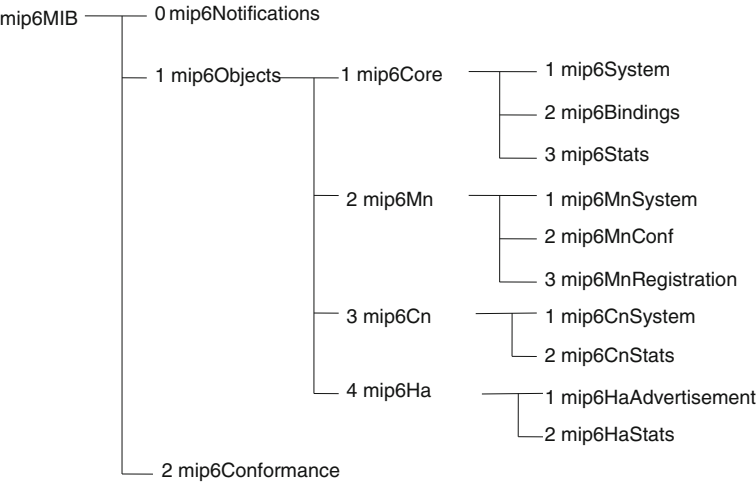


Fig. 2.3 Structural description of MIB definition for MIPv6. Reprinted from IETF RFC4295 [4]

(2) Network management protocol and MIB

Taking Simple Network Management Protocol (SNMP) as an example, it is one of the most commonly used network management protocols. SNMP was defined by the IETF and is expected to realize simple but efficient network management. The management information base (MIB) is the database that records all kinds of information about the managed objects. The MIB extension for MIPv6 is defined by RFC4295 [4], which can be used to monitor and control the MIPv6 entities such as the MN, the HA, and the CN. The detailed monitored information includes functions of MIPv6 entities, MIPv6 service traffic, binding data, and the update history at the MN, CN, and HA.

A structural description of the MIB definition for MIPv6 is defined in [4]. As a part of the whole management information base, the structural description of the mip6Notifications group is shown in Fig. 2.3.

2.2 Network Reference Model

According to the above protocol reference model, the network reference model, and the functional entities of mobility management are abstracted, as shown in Fig. 2.4. Depending on where the mobile object is subscribed, the network is divided into one home network and multiple visiting (foreign) networks. The home network and the visiting networks may then be further divided into subareas (e.g., Location Areas (LAs) and Paging Areas (PAs) in cellular systems). There are four function entities: mobility management servers (including the corresponding databases),

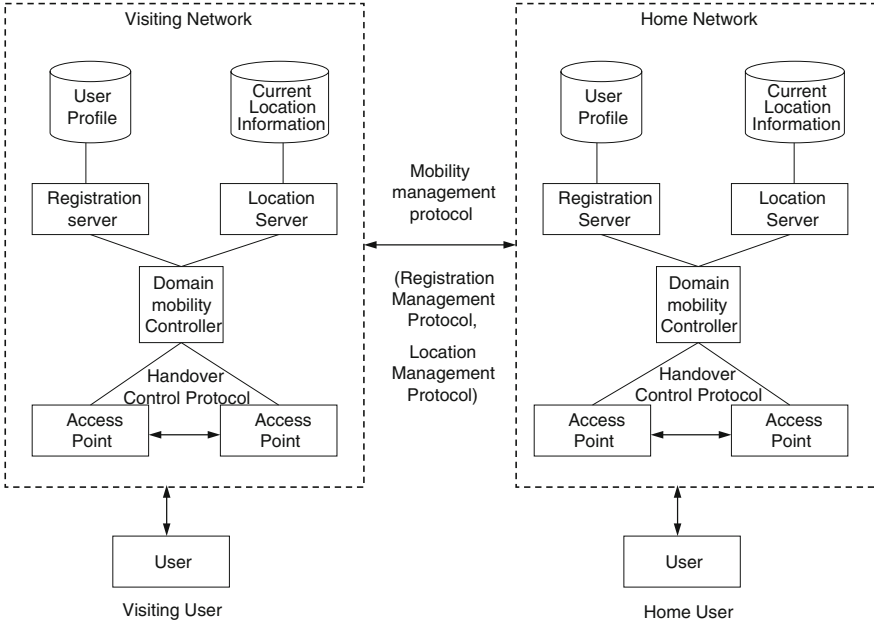


Fig. 2.4 Mobility management network reference model and functional entities. Reprinted from ref. [1], Copyright 2007, with permission from Journal of communications

network access points, the domain mobility controllers, and the corresponding mobility management protocols.

2.2.1 Mobility Management Server

The mobility management servers are responsible for providing the mobility management functions, including the registration servers and the location servers.

The registration servers maintain the users' access authentication information (including network access authentication and service access authentication) and service profile information, but they may be separated into different entities during implementation.

Taking the GSM network as an example, the home location register (HLR) and the authentication center (AuC) cooperate to act as the registration server. HLR records the local user information such as user identity, registration information, roaming capability, subscribed services, and supplementary services. AuC is used to store root key (K_i) and generate a triple [randomly generated number (RAND), signed response (SRES) and the cryptographic key (K_c)] needed to authenticate the user and provide ciphering protection. The HLR stores the triples for each user and transmits them to the Mobile Switching Center (MSC)/Visitor Location Register

(VLR) when required. The MSC/VLR sends RAND to the mobile station, and the mobile station uses RAND, the root key K_i , and the same algorithm as the AuC to compute SRES. The resulting SRES is sent back to MSC/VLR for authentication.

Another example is the Hypertext Transfer Protocol (HTTP)-based authentication mechanism in SIP. It adopts a stateless, challenge-response-based authentication. The challenge and the credentials are exchanged between the user agent client (UAC) and the different SIP servers.

The location servers are responsible for recording, updating, finding, and deregistering the users' current location information. The registration servers and location servers in a visiting network only store the temporary information of any visiting users roaming to their network. When the user roams out of this visiting network, the corresponding information is deleted. The registration server in the home network stores the users' permanent (or semipermanent) information. The location server in the home network also communicates with the location server in the visiting network to update the current user location information.

In MIPv4, the HA and FA behave as the location servers in the home network and the visiting network, respectively. The FA maintains the current location information (i.e., care-of address (CoA)) and the data are routed to the MN accordingly. The HA maintains the home of address (HoA) along with the binding of the HoA with the CoA. A registration request message in MIPv4 is used to update the binding.

2.2.2 Access Point

The access points provide functions including communication access, number/address mapping, and the corresponding handover control and location management related functions. The base station subsystem (BSS) in cellular networks and the AP in WLAN are examples of access points. Handover control will occur between adjacent access points, when the user movement leads to a change in the access points.

The GSM network is used as an example again. Acting as the access point, the BSS also takes part in the handover and location management along with its wireless access function. If the handover target cell is within the same BSS, the handover will be controlled by the base station controller (BSC). Otherwise, the handover request will be sent to the target BSC through the MSC. For location retrieval, the BSC is responsible for paging within all the cells under its coverage. For location updates, when the mobile station moves into the coverage of a BSS in another MSC, it will detect the change in wireless signal quality and the location area code (LAC). The current BSS and MSC will then send a location update request to the new VLR. Subsequent location operations will then be completed by the new VLR and HLR.

2.2.3 Domain Mobility Controller

The network is partitioned into multiple domains for convenience in mobility management. The mobility across different domains is handled by the domain mobility controller. The MSC/gateway MSC (GMSC) in cellular networks, and the HA and FA in MIP are all examples of domain mobility controllers. The domain mobility controller takes responsibility for: (1) handover control between the different access points within a single domain, and across the domains between different visiting networks; (2) the domain gateway (i.e., the exchange of registration and location information between the home network and the visiting network).

The definition of a domain may vary between different mobility management technologies. In GSM networks, the coverage of an individual MSC, or the cell groups under the same BSC, could be defined as a domain. The MSC is in charge of inter-MSC handovers and inter-BSC handovers within the same MSC. The MSC also takes part in the location update and paging functions.

Here, we use the MIPv4 network as another example. HAs and FAs can be considered as the domain mobility controllers of the subnets where they are located. When the MN leaves its home network, the HA is responsible for forwarding the data packets for the MN to the FA through tunnels. After the FA receives the data packets, it depacketizes the data and then sends them to the MN. Thus, the handover procedure is complete. The HA and FA also exchange messages with the MN for location registration.

2.2.4 Mobility Management Protocol

The mobility management protocols are responsible for communications between the functional entities to realize authentication, information exchange (e.g., user information and location information), and control functions. There are four types of protocols: the registration management protocol, the location management protocol, the handover control protocol, and the interoperability protocol. For example, the registration management protocol is responsible for transmission and copying of the service profile data from the home network to the visiting network to guarantee the consistency of the user service environment.

The GSM network is again a good example. In GSM networks, there are typically three types of mobility management protocols: (1) the Mobile Application Part (MAP) of Signaling System No. 7 (SS7), which is used for communications between the MSC, HLR and VLR and is involved in registration management and location management; (2) the Base Station System Application Part (BSSAP), which is used for signaling between the MSC and the BSS and is involved in handover control; and (3) Um interface signaling, which is used for signaling between the mobile station and the BS and is involved in paging and location registration.

Taking the MIPv4 network as another example, its mobility management protocols include agent discovery and Registration Request messages. Agent discovery is realized through agent advertisement messages, which are periodically multi-cast or broadcast by the HAs and FAs. A registration request message is used by an MN to register its new CoA to the HA. The binding of the HoA and the CoA can be set in the HA for location management. For MIPv6, mobility detection, and binding update messages are used for similar functions.

2.3 Critical Control Functions in Mobility Management

2.3.1 Security Mechanism

The threats in mobile environment exist in many aspects such as loss/theft of mobile devices, data interception and tampering, malware, vulnerable applications, compromised devices, Web browser exploration, and OS vulnerability [5, 6]. Security mechanism in mobility management handles the following functions:

(1) Critical data privacy protection

The critical data in mobility management include the user identification information, the user location information, and the signaling messages involved in mobility management. Therefore, two major functions are necessary to provide critical data privacy: (A) To provide the privacy of user identification information and location information; (B) To provide the privacy of signaling messages exchanged between different entities participating in mobility management. The privacy of user data is not included here.

(2) Registration and authentication management

This function is responsible for the management of user registration information and service profile information. AAA is used for: authentication and authorization for network access of the mobile users and mobile terminals; authentication and authorization for service access of the mobile users and mobile terminals; accounting to track the user's utilization of network resources for analysis, audit, and billing. Service profile management handles the registration, update, lookup, and deregistration of the service profile information.

(3) Signaling message integrity

This function is used to provide integrity for the various signaling messages involved in mobility management to keep them from being tampered.

(4) Mobile service non-repudiation

This function is used to keep the mobile objects from fraud of the internal legitimate users in the mobility management system, rather than to prevent threats

from external unknown attackers. Mobile service non-repudiation means both service source and service data receipt are undeniable.

2.3.2 Location Management

In mobility scenarios, the mobile objects (mobile users or mobile terminals) frequently move from one place to another. To track the location of the mobile object for effective mobility support, the location information must be stored in some specific entities (e.g., the location server) in the network and be retrieved when required. As one of the main control functions of mobility management, location management is responsible for storing, updating, and retrieving the locations of the mobile objects.

(1) Functions of location management

There are two important functions in location management: location update (also called location registration) and location retrieval (also called paging).

Location update refers to the event where the mobile object notifies the system of a location change such as the location area (LA) update in 2G cellular systems. When a location update occurs, the system will update the location database and determine the routing for the newly arrived call or data accordingly. The important points in location update are how the mobile object detects the mobility, and when and how to report the object's current location. **Paging** refers to the procedure in which the system finds the locations of mobile objects, with the aim of determining the locations of the objects effectively and precisely, such as the paging function in 2G systems.

Various methods have been proposed for effective location management. In fact, analysis of the extreme conditions in location information storage and location update policy is helpful in understanding the trade-off between location update and paging.

For location information storage, there are two extreme conditions [7]. In condition (a), the location information is stored at each entity in the network, thus making it easy to acquire location information with low paging overhead. However, the location update overhead is extremely high in this case, because every database maintained at each of the different entities must be updated when a mobile object changes its location. In condition (b), the location information is not stored in any of the network entities. Thus, location of a mobile object requires a search of the whole network (e.g., paging through broadcasting). The paging overhead is therefore extremely high, while the location update overhead is very low.

Two extreme conditions also exist for the location update policy [8]. Condition (a) is the "always update" policy, where the location database is updated when any mobile object changes its location, which incurs frequent location update operations and the associated high overhead. However, paging at this time becomes very easy

because the locations recorded in the database always contain the most accurate and latest location information. Condition (b) is the “never update” policy, in which the location database is never updated, even when the mobile object changes its location. This policy incurs an extremely high paging overhead but without any location update overhead.

Based on the above analysis, it is easy to understand the contradiction in the system resource usage of location update and paging. More frequent location updates can improve the paging efficiency but incurs correspondingly greater signaling overhead. Research in location management schemes is aimed at achieving a trade-off between the two.

In general, the various location management policies try to find a balance among availability, precision, and currency [9]. Availability relates to the location information storage policy, ranging from storing location information at every network entity to not storing location information at any of the network entities. Precision relates to the recorded location information content (i.e., whether the location is a possible location or a precise location). Currency relates to when the location information should be updated. For example, for mobile objects with low mobility but with high incoming call probability, the location update need not be performed each time when the object moves.

(2) Databases in location management

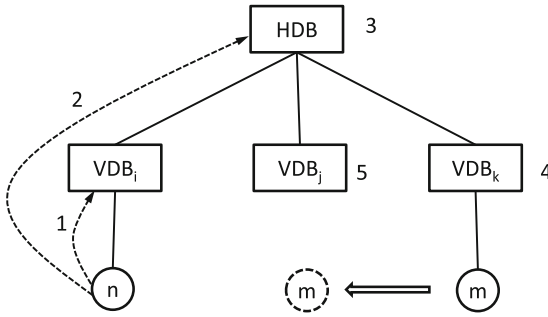
Location management requires a location database. The HLR and the VLR in 2G systems, the HA in the MIP system and the location server in the SIP system are all examples of these databases.

The location database may use several different database structures.

A. Hierarchical database

In the hierarchical database, the location information of the mobile objects is recorded in the databases of different hierarchies. A two-level database structure is normally used, in which two types of databases, the home database (HDB) and the visiting database (VDB), are deployed [7]. Every network is configured with one HDB and multiple VDBs. The HDB retains the registered user-related information (e.g., the user ID, access authority, password, and current location). Each user is associated with one HDB. The VDB holds the related data of the users resident in the associated location area. The data in the VDB are often part of the data in the HDB. The HLR and the VLR in the 2G system, the HA and the FA in MIPv4, and the home location function (HLF) and the visitor location function (VLF) in H.323 could all be considered as typical examples of two-level databases.

The basic location operations of two-level databases can be summarized as follows. (i) location retrieval: if a user in LA i calls the mobile user m , then the location retrieval operation will be conducted first in the VDB of LA i . The operation will then continue in the HDB of user m only if m 's data cannot be found in the VDB of LA i . (ii) Location update: when user m moves from LA k to LA j , m 's location data in his HDB will be updated. At the same time, the m -related data



Location retrieval: 1-2 in this figure describe the operation when user n in LA i calls user m in LA k .
 1: location retrieving firstly in VDB_i ;
 2: location retrieving is continued in HDB of user m when m 's data can't be found in VDB_i ;

Location update: 3-5 in this figure describe the operation when user m moves from LA k to LA j .
 3: m 's location data in his HDB is updated;
 4: m 's location data in VDB_k is deleted;
 5: m 's location data is added in VDB_j .

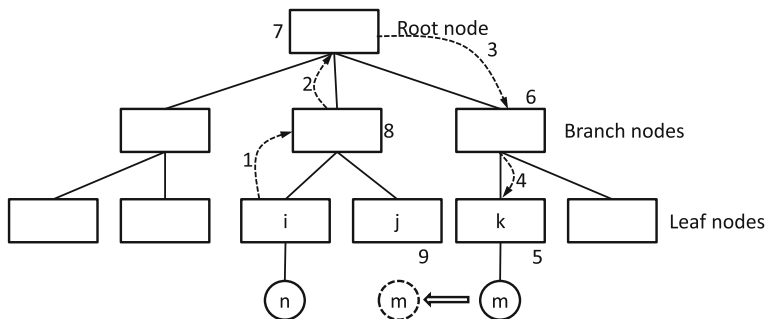
Fig. 2.5 Location operations in a hierarchical database

will be deleted from the VDB of LA k , while being added to the VDB of LA j [7]. Figure 2.5 illustrates the location operations in a hierarchical database.

For a two-level location database, the location of the HDB does not change generally. The location retrieval operation often begins in the HDB. Therefore, for a user that is always resident in distant foreign networks, long-distance communication is required to access their associated HDB whenever they roam from one visiting LA to another adjacent visiting LA. This type of operation cannot use the chain-like characteristics of the multiple adjacent LAs and thus incurs heavy overheads and bandwidth consumption. This is the primary disadvantage of the two-level location database [7].

B. Tree structure database

In the tree structure database, the tree is used to describe the network locations, and the location information of a mobile object is maintained in the corresponding leaf node and branch node. An upended tree is used in the tree structure database. Each leaf node is associated with an individual LA, and a location database is deployed at the leaf node for this LA, recording the location information of all users resident in it. The database is also deployed at the branch node and at the root node to record the location information of users in all the LAs belonging to its subtree. Also, a pointer exists in the database at every non-leaf node, pointing to the database of its child nodes [7].



Location retrieval: 1-4 in this figure describe the operation when user n in LA i calls user m in LA k .

- 1-2: searching m 's location data from i upward to find LCA (i, k);
- 3-4: retrieving operations from the LCA (i, k) downward to k .

Location update: 5-9 in this figure describe the operation when user m moves from LA k to LA j .

- 5-6: m 's location data in related nodes are deleted;
- 7: m 's location data in LCA (k, j) (i.e., the database pointer) is updated;
- 8-9: m 's location data is added in related nodes.

Fig. 2.6 Location operations in the tree structure database

The basic location operations of the tree structure database can be summarized as follows. (i) Location retrieval: if the mobile user in LA i calls mobile user m in LA k , then the location retrieval operation will look for the node containing the location information of user m from node i upward. Obviously, it is simply the least common ancestor (LCA) node of i and k . The querying then goes from this LCA node to node k and user m will be found. (ii) Location update: if the mobile user m moves from LA k to LA j , then the location databases along the path from node k upward to the LCA node of k and j , and then downward to node j will all be updated [7]. Figure 2.6 illustrates the location operations in the tree structure database.

For the scenario mentioned above, in which the user moves between the adjacent LAs frequently, the tree structure database can decrease the communication overhead greatly because it can operate without any need for long-distance communication and only requires access to several adjacent databases. This is the apparent advantage of the tree structure compared with the two-level database. However, the disadvantage of the tree structure database is the large management overhead because of the large number of databases. It also has more stringent performance requirements in terms of capacity and speed for the upper-level nodes [7].

C. Centralized database

In the centralized database, the current location information for all mobile users is maintained in a single, centralized database [7], rather than the two-level database and the tree structure database. Obviously, the advantages of the centralized

database include its simple structure, easy maintenance, simple location operations, and easy data consistency and integrity. However, it also suffers from the common problems of a centralized system, such as scalability and high reliability, stability, capacity, and access speed requirements for the centralized database system.

(3) Analysis of the location management overhead

In location management, the entire network is often partitioned into multiple LAs and paging areas (PAs). The LA is the basic unit for location tracking. When the mobile terminals move within the area of the same LA, it requires no location update operations. However, when a mobile terminal moves across the border of the LA, it should update its location information in the system. In cellular systems, the LA is usually composed of multiple cells. The coverage and shape of the LA could be constant or could change dynamically according to the actual network environments. The PA is the area within which the system searches for the mobile terminal via broadcasting. Depending on the different paging policies, the PA could be limited to a single cell or composed of all the cells in the system.

In future mobile communication systems, tremendous numbers of mobile users will move at random in a large area (even globally) and may communicate at any moment. Therefore, the deployment of the location servers and the planning of the LAs and PAs are very important for fast and precise callee location finding. The planning and optimization of the LAs and PAs are correlative with each other. Depending on the different targets and policies of LAs and PAs, a PA's coverage could be smaller than or equal to that of an LA's. The larger the LA is, the lower the location update frequency becomes, and thus the smaller the update signaling overhead becomes, but the larger that a PA is, the larger the paging signaling overhead becomes.

Performance evaluation parameters for location management include the update signaling overhead, the update delay, the paging signaling overhead, the paging delay, and the system efficiency.

The total cost function for location management can be described as follows:

$$C_{LA} = C_{LU} + C_{PAG}$$

where C_{LA} is the total cost of location management, C_{LU} is the location update cost (including the update signaling overhead and the update delay), and C_{PAG} is the cost of paging (including the paging signaling overhead and the paging delay). The most optimized LA and PA planning schemes aim to minimize this total cost function. In fact, location update and paging are in conflict with each other in terms of system resource consumption, and thus the various location management schemes try to achieve a trade-off between them. Table 2.3 compares the location update and location retrieval and gives an analysis of their overheads.

(4) Location management policies

To achieve the trade-off between the location update and location retrieval, various improved location management schemes have been proposed. According to

Table 2.3 Comparison of location update and location retrieval

Function	Originator	Overhead	Overhead versus LA and PA coverage
Location update	Mobile object	Incurred by update signaling, which consumes the wireless channel resources and increases the load and processing delay of the network and location databases	The larger the LA is, the lower the update frequency becomes, and thus the lower the overhead becomes
Location retrieval	Network	Incurred by paging signaling, which consumes the wireless resources	The smaller the PA is, the lower the paging overhead becomes, and thus the higher the paging success rate becomes with lower paging latency

the update policy of the location database, these location management schemes can be classified into two types: static policy-based schemes and dynamic policy-based schemes. Static policies include the basic policy, the pointer extending policy, the anchor-based policy, the anchor- and pointer-integrated policy, and the circle searching policy. Dynamic policies include the time-based location management policy, the mobility-based location management policy, and the distance-based location management policy. More details of these different policies can be found in reference [7].

Further, most existing mobility management schemes are considered to be passive, because they passively track the location changes of the mobile objects and maintain the network connectivity for them. This passive feature means that these schemes are unlikely to satisfy the requirements of future mobile users. Therefore, some predictive mobility management schemes were proposed. These schemes model the mobility patterns and use prediction algorithms for effective prediction of the possible location of the mobile users. This is the basis of predictive mobility management.

2.3.3 Handover Control

Handover control is used to maintain session continuity for the user when the access point changes during movement (i.e., the ongoing communication provided by the current access point is transferred to another access point rather than being interrupted). As one of the critical control functions and technologies in mobility management, handover control has attracted considerable research effort for both traditional mobile communications and future ubiquitous and heterogeneous network environments.

(1) **Classification of handover**

Depending on different criteria, we can obtain diverse handover types, as shown in Table 2.4.

- (A) Based on the mobility range, handover could be classified into intranetwork handover and internetwork handover. We can see examples of such classification in different networks, such as intra-access service network (intra-ASN) handover and inter-ASN handover in worldwide interoperability for microwave access (WiMAX) networks. Also, in 2G systems, intranetwork handover could be further subdivided into intracell handover, intrabase station controller (BSC) handover, intramobile switching center (MSC) handover, and inter-MSC handover. In IP networks, intranetwork handover can be subdivided into intra-access router (AR) handover, intra-access network (AN) handover, and inter-AN handover.
- (B) Depending on the homogeneity or heterogeneity of the network types involved, handover could be classified into horizontal handover and vertical handover [10, 11]. The difference is that horizontal handover is performed between different wireless access points that use the same technology, while vertical handover involves changing the access technology. For example, handover within the Global System for Mobile Communication (GSM) network is a horizontal handover, but a handover between a Universal Mobile Telecommunication System (UMTS) network and a WLAN is a vertical handover.

Table 2.4 Handover classifications

Classification criteria	Classifications
Mobility range	Intranetwork handover
	Internetwork handover
Network types involved (homogeneous or heterogeneous)	Horizontal handover
	Vertical handover
Handover performance requirements	Fast handover
	Smooth handover
	Seamless handover
Handover purpose	Rescue handover
	Confinement handover
	Traffic handover
Handover procedure	Hard handover
	Soft handover
	Softer handover
Handover necessity	Forced handover
	Unforced handover
User control (permitted or not permitted)	Active handover
	Passive handover

- (C) Based on the handover performance requirements, handover could be classified into fast handover, smooth handover, and seamless handover [12]. Smooth handover aims to minimize the packet loss during the handover procedure. Fast handover aims to minimize the handover delay. Seamless handover requires high performance in both packet loss and handover delay. This is the handover type that means no change in user experience in terms of service capability, security, and quality of service (QoS).
- (D) Based on different purposes, there are rescue handover, confinement handover, and traffic handover [13]. Rescue handover is handover when the mobile station leaves the area covered by one cell and moves into another area. It is the quality of the transmission that determines the handover necessity, where the quality is indicated by the error rate, the received signal strength, the interference level, and the propagation delay. Confinement handover is the handover performed when the mobile station would suffer less interference if it changed cell (interference is caused in part by other mobile stations in the cell). The mobile station continuously listens to other cells to measure the quality of a connection to the latter. Each mobile station is also synchronized with several base transceiver stations (BTSs) to be ready in case of handover. Traffic handover occurs when the number of mobile stations is too large for a cell and the neighboring cells can accommodate new mobile stations. This decision requires knowledge of the charge of other BTSs.
- (E) Depending on the handover procedure (i.e., whether the old link is released before or after the new link is established), we have hard handover, soft handover, and softer handover [14]. This classification originates from the handover technology in traditional cellular systems. A hard handover is one in which the channel in the source cell is released first, and only then is the channel in the target cell engaged. The connection to the source is thus broken before or 'as' the connection to the target is made; for this reason, these handovers are also known as break-before-make. In traditional mobile communications, the handover between base stations (BSs) or sectors using different frequencies could only adopt hard handover. This process could also be used for handover between BSs or sectors belonging to different operators or different systems. A soft handover is the one in which the channel in the source cell is retained and is used for a while in parallel with the channel in the target cell. In this case, the connection to the target cell is established before the connection to the source cell is broken, and thus, this handover is called make-before-break. Soft handover is implemented through the control of the MSCs. It is only used for handover between different BSs using the same frequency. Soft handovers may involve using connections to more than two cells: connections to three, four, or more cells can be maintained by a single mobile station at the same time. When a call is in a soft handover state, the signal that is the best of all used channels can be used for the call at any given moment, or all the signals can be combined to produce a clearer version of the signal. The latter is more advantageous, and when such signal combining is performed in both the downlink (the forward link) and the uplink (the reverse

link), the handover is termed a softer handover. Softer handovers are possible when the cells involved in the handovers have a single cell site. In general mobility management for heterogeneous networks, the hard handover and soft handover are no longer limited to the cellular communication field. They have been extended to various mobility management technologies. For example, soft handover schemes were proposed for MIP and SIP improvement to lower the packet losses during handover for handover performance optimization.

- (F) Depending on the handover necessity, handover could be classified into forced handover and unforced handover [15]. Forced handover is often triggered by the events regarding network interface availability. It occurs when there is only one available interface left. Handover is necessary here to avoid communication interruption. Unforced handover usually happens when multiple interfaces are available simultaneously. The target of unforced handover is to improve the QoS.
- (G) Depending on user control allowance, we also have active handover and passive handover [16]. Active handover allows the user to participate in handover control; the handover is based on the user preference.

(2) **The major technical issues of handover control**

The major technical issues in handover control are the handover rule, the handover control mode, resource allocation during the handover, and communication link transfer.

A. **Handover rule**

The handover rule determines when and under what conditions the handover actions are triggered. In traditional handover for cellular systems, the handover rules are generally determined based on the measured wireless signal strength and the channel quality [17]. For vertical handover across heterogeneous access networks, the handover decision should involve more factors with respect to the user, the application, the network, and the terminal.

Figure 2.7 shows an example of the handover rule in a traditional cellular system—the handover rule for soft handover in a code division multiple access (CDMA) system [17]. It illustrates the variation in the pilot signal strength of a mobile station during soft handover from base station A to base station B. The mobile station measures the pilot signal strength of the adjacent cells while it maintains communication through BS A. When the signal strength of one cell (e.g., BS B in this example) becomes higher than a certain threshold (the upper limit in Fig. 2.7), it will report to the system. The system then tells BS B to establish communication with the mobile terminal and the soft handover is triggered. At this time, the mobile station receives signals from both A and B in parallel. When the mobile station detects that the signal strength of the source cell (BS A in this example) has fallen below another threshold (the lower limit in Fig. 2.7), it starts a timer. If the signal strength of the source cell remains lower than the threshold for a predetermined period, then the mobile station will disconnect from the source cell. The soft handover procedure then terminates.

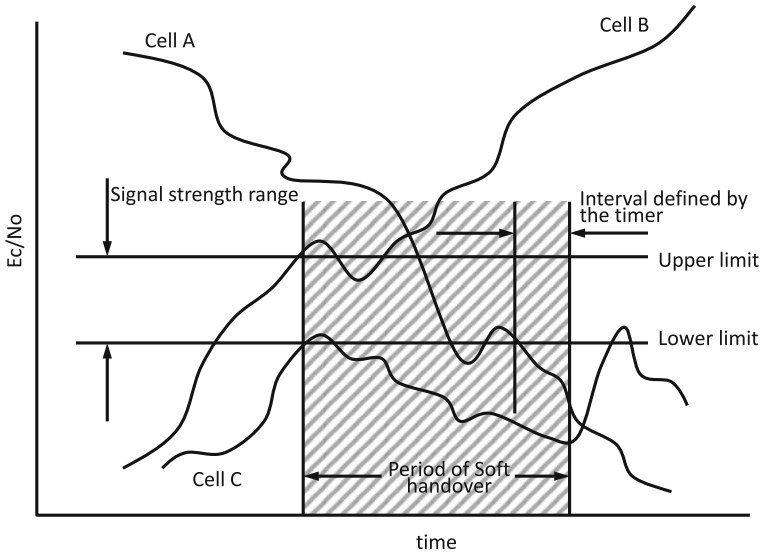


Fig. 2.7 Handover rule for soft handover in CDMA system. Reprinted from ref. [17], Copyright 1999, with permission from the author

The vertical handover decision is a typical multi-criteria decision problem. The decision is made based on various network-related, application-related, and user-related factors. Network-related factors include access technology, coverage range, signal strength, available bandwidth, error rate, actual delay, and other network features and status information. Application factors describe the current application features and requirements. User-related factors include the user's preferences, terminal capabilities, and user mobility characteristics. More details regarding the vertical handover decision will be presented in Sect. 3.1.2.

B. Handover control mode

The handover control mode includes the handover detection and control strategies. It determines who is responsible for handover-related information collection (e.g., measurement of the wireless channel quality), how to collect and report this information, and who initiates the handover.

Based on the roles of the network and the mobile terminals, there are three commonly used handover control modes: network-controlled handover (NCHO), mobile-assisted handover (MAHO), and mobile-controlled handover (MCHO) [18].

In the NCHO mode, the surrounding access points (e.g., the BSs) measure the signal from the mobile station. The network initiates the handover process when certain handover criteria are met. The first-generation analog cellular communication systems, including the Total Access Communication System (TACS) and the Advanced Mobile Phone System (AMPS), used the NCHO mode, where the MSC

controls the handover by following a centralized control mode. NCHO supports intercell handover only.

In the MAHO mode, the network asks the mobile station to measure the signals from the surrounding access points (e.g., the BSs). The network makes the handover decision based on the reports from the mobile station. 2G cellular communication systems such as GSM and CDMA adopted the MAHO mode. Generally, the MAHO mode could support both intercell and intracell handover.

In the MCHO mode, the mobile station continuously monitors the signals of the surrounding access points (e.g., the BSs). The mobile station initiates the handover process when certain handover criteria are met. Systems including WLAN, Digital European Cordless Telecommunications (DECT), and Personal Access Communications System (PACS) use the MCHO mode. The MCHO mode is also capable of supporting intercell and intracell handover.

C. Resource allocation during handover

Resource allocation handles the related resource assignment during handover (e.g., channel assignment in cellular systems, CoA assignment, and IP address binding in MIP systems). Taking the cellular system as an example, the handover call will be terminated (known as forced termination) if there is no available channel on the selected access point. If non-prioritized scheme (NPS) is adopted, a handover call and a new arrival call will be treated in the same way. If there is no available channel, the handover call will be blocked and remains on the original channel until the call is complete or the channel becomes unavailable. Such a scheme may incur forced termination, which is regarded as degradation of the service quality because forced termination is considered to be more serious than new call blocking in terms of the user experience.

Some trade-offs among service quality, spectrum efficiency, and implementation complexity should actually be considered as part of the channel assignment algorithm design [19, 20]. A good algorithm is expected to achieve high spectrum efficiency and low complexity while achieving the required service quality (link quality, new call block rate, and forced termination rate). In this respect, other schemes including reserved channel scheme (RCS), queuing priority scheme (QPS), and subrating scheme (SRS) are considered to be better than NPS.

In RCS, a number of channels in each BS are preserved for handover calls, while the other operations are similar to those of NPS. In other words, the channels are divided into two groups: preserved channels that serve only handover calls, and other channels that serve both new arrival calls and handover calls.

In QPS, a handover call is treated with a higher priority than new calls, i.e., it is easier for a handover call to be allocated a channel. QPS is based on the fact that the adjacent cells in a PCS network overlap each other. Thus, there is a considerable area where a call can be handled by the BS in either of the adjacent cells, which is called the handover area. The time that a mobile station spends in the overlapped area is referred to as the degradation interval. If no channel is available for the handover call after the mobile station moves out of the handover area, i.e., when the

degradation interval expires, the call is forced to terminate. In this scheme, when a channel is released, the BS first checks whether the waiting queue is empty. If not, the released channel is assigned to a handover call in the queue. The next handover to be served is selected based on the queuing policy, for example, a simple first-in first-out (FIFO) policy and a measure-based priority scheme (MBPS) [19].

SRS creates a new channel on a blocked BS for a handover access attempt by subrating an existing call. Subrating is the process of temporarily dividing an occupied full-rate channel into two channels at half the original rate; one serves the existing call while the other serves the handover request [20].

D. Communication link transfer

Communication link transfer is the procedure in which the communication link is transferred from the existing access point to the target access point. The network must bridge the link to the new access point and then drop the link to the old access point.

The link transfer may be made from one channel to another channel on the same access point, for example, intracell handover in the time division multiple access (TDMA) system, or from one access point to another access point attached to the same network control entity, such as inter-BS handover under the same BSC, or inter-BSC handover under the same MSC. The link may also be between access points belonging to different systems (e.g., handover between heterogeneous access networks). Based on the link transfer strategy, we have the hard handover and soft handover classifications.

The handover control function in the cellular mobile communication system is a mature technology. As shown above, it is used as a good example in the introduction of the major functions of the handover process. It should be noted that handover control functions in future heterogeneous network environments are likely to be different from those in cellular systems. Table 2.5 compares the handover control functions in cellular systems with those for heterogeneous networks.

(3) Performance evaluation of handover algorithm

An excellent handover scheme could maximize the system capacity. However, there are several other handover performance evaluation parameters, including the handover success rate, the call drop rate, the new arrival call blocking rate, the average handover frequency, the handover delay, and the forced termination rate.

The handover success rate is the percentage of successful handovers relative to the total number of handover attempts. The call drop rate is the ratio of the number of dropped calls to the total number of successful calls, where “the total number of successful calls” includes both newly initiated calls and handover calls. The new arrival call blocking rate is the ratio of the number of blocked new calls because of wireless resource shortages to the total number of new calls. It should be noted that this parameter does not involve the call blocking incurred for other reasons (e.g., in the shadow area). The average handover frequency is the average handover time throughout the entire session. The handover delay refers to the communication

Table 2.5 Comparison of handover control functions in cellular system with those in heterogeneous networks

Handover control function	Handover in cellular system	Handover in heterogeneous networks
Handover rules	Mainly based on measurements on wireless links; Commonly used measurement methods for link quality: WEI (word error indicator), QI (quality indicator) and RSSI (received signal strength indicator)	Multi-criteria decision problem involving user-related, application-related, network-related, and terminal-related factors
Handover control mode	MCHO, NCHO and MAHO	Commonly used control modes: network initiated, mobile terminal initiated, network and mobile terminal co-initiated
Resource allocation	Mainly refers to wireless channel assignment, including NPS, RCS, QPS, and SRS	For example, CoA assignment and IP address binding in MIP
Communication link transfer	Executed at data link layer	Executed at network layer, transport layer, or application layer

interrupt period caused by the handover. The forced termination rate is the ratio of the calls that were successfully initiated but were then terminated because of handover failure to the total number of calls.

2.3.4 Interoperability Control

Given the diversity of available access technologies and the abundant mobility objects and types, interoperability control is the control function required for mobility management in heterogeneous networks. Here, the interoperability control function involves two aspects: (1) handling the gap between the different access technologies for mobility-related issues; (2) cross-layer methodology in mobility management.

(1) Handling the gaps between different access technologies for mobility-related issues

Users with multiple interfaces (using multi-interface terminals or multiple heterogeneous terminals) while moving across heterogeneous access networks will be the prevalent scenario in the future. In this case, the user is located within the overlapped area covered by heterogeneous access technologies. In such a scenario, the users expect to enjoy their subscribed service seamlessly, and thus, the heterogeneous access networks are required to cooperate with each other to provide the best possible service experience based on their complementary features.

Therefore, interoperability control is used here to handle the gaps between the different access technologies for mobility-related issues.

The best network selection could be considered as an example of interoperability control. A user located within the overlapping area of multiple heterogeneous access networks could connect to the networks via different network interfaces. Selection of the best network for both new services and handover services is an important issue for the best service experience. Here, the best network selection should be the result of a decision based on multiple factors about the networks, users, and applications. The AAA mechanism across heterogeneous networks is another issue in interoperability control. A new AAA mechanism should be designed to be able to adapt flexibly to multi-network or multi-operator environments, to ensure security, and to provide a billing mechanism for intercarrier roaming users. Another inevitable interoperability control issue is QoS adaptation across the heterogeneous access technologies. QoS control mechanisms are defined individually in different wireless access technologies. For example, a cellular system has a strict QoS guarantee mechanism, and wireless broadband access networks like WLAN and WiMAX have also defined their own QoS control schemes. When the user moves across these heterogeneous access technologies, the differences in the QoS control mechanisms and the unpredictable nature of the wireless link quality consequently result in obvious fluctuations in QoS. In this case, a QoS adaptation mechanism is required for seamless mobility support in heterogeneous network environments.

(2) Cross-layer methodology in mobility support

The cross-layer methodology requirements in mobility management stem from two different aspects. On the one hand, mobility management technology at an individual protocol layer could not meet the abundant functionality and performance requirements. Application of cross-layer design is a good choice for integration of these technologies for complementation. On the other hand, the traditional protocol layer design strictly follows the hierarchical layered structure. However, in wireless and mobile network environments, serious performance degradation will be introduced by wireless channel quality instability, scarcity of radio resources and variation in the wireless link features caused by mobility across the different access technologies. A cross-layer design methodology is therefore required for coordination and cooperation between the various technologies at the different layers to achieve better mobility management functionality and performance.

Applications of cross-layer design in mobility management include:

A. Cross-layer integration

Cross-layer integration is based on the complementarity of the various mobility management technologies belonging to the different layers. This cross-layer integration is applied to the network entities and to the signaling design.

Cross-layer integration of the MIP at the network layer and the SIP at the application layer is a good example of this type [21–23]. The complementary capabilities of MIP and SIP can be summarized as the following [21]: (i) The MIP provides transparent mobility to the upper layers and hides the changes in IP address to the application layer of the MN and the correspondent node (CN). The SIP provides the application layer mobility. (ii) The MIP is usually based on TCP, while the SIP is often based on the User Datagram Protocol (UDP); (iii) The MIP is often used to support terminal mobility, while the SIP is used to support high-level mobility such as personal mobility and service mobility.

Accordingly, the integration of MIP and SIP could be used in the design of network entities. Both MIP and SIP require the deployment of some servers in the network, for example, the HA (and the FA in MIPv4) in MIP and the registrar, the location server, the redirect server and the proxy server in SIP. For the integrated scheme, similar functions in different protocols could be integrated at suitable network entities [21]. The MIP/SIP integration could also be used in the design of signaling procedures [21, 23]. Signaling procedures exist for similar functions in MIP and SIP. For example, the MN will execute a location update operation in both MIP and SIP when its location changes. The MN registers its new location information with different entities (HA in MIP and the home registrar in SIP), carrying similar information in the registration messages. This procedure of “moving once, registering twice” brings serious redundancy functionality and excess signaling. This will result in low system efficiency, high packet loss, and handover delay. Therefore, cross-layer integration should be applied to redundant function and signaling optimization.

Another example occurs at the network layer and the transport layer (i.e., the cross-layer integration of mSCTP and MIP). mSCTP provides end-to-end mobility at the transport layer and has a handover control function but does not have a location management function. Integration of mSCTP and MIP would expect to use MIP for location management (i.e., locating the mobile terminal that is being called), and then, the mSCTP association between the two endpoints can be established. After that, mSCTP is used to provide handover for the communication endpoints [24].

B. Cross-layer information interaction

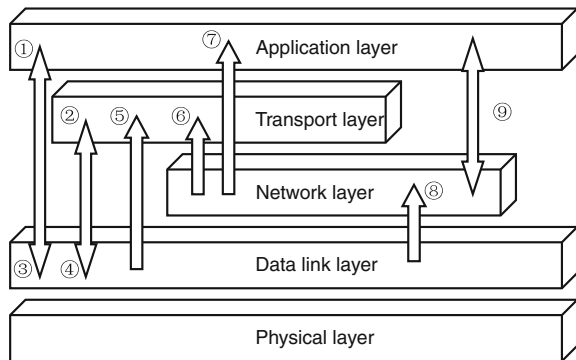
Cross-layer information interaction is often used for mobility management, and especially for handover performance optimization [25]. Figure 2.8 illustrates the possible cross-layer information interactions at the mobile node.

- ①② The link layer conveys the link actions (e.g., frame loss and delay) and link status (e.g., current available bit rate) to the application layer and the transport layer. Then, the application can differentiate the various cases for different handling, thus avoiding the performance degradation that results from these cases simply being treated as congestion.

- ③④ The application layer and the transport layer inform the data link layer about the delay requirements and the tolerable error rate, so that the data link layer can select suitable modulation, error correction, and check mechanisms accordingly.
- ⑤⑥ The data link layer and the network layer notify the TCP at the transport layer about the handover commencement and handover completion events. The TCP can then adjust its congestion control mechanism accordingly, including the congestion window (CWND), the round trip time (RTT), the retransmission timeout (RTO) and the congestion control algorithms (e.g., the slow start algorithm and the congestion avoidance algorithm), for TCP performance optimization.
- ⑦ The application layer adjusts the data sending rate according to the handover notification from the MIP to reduce packet losses during handover and optimize the handover performance.
- ⑧ Data link-layer information is provided for MIP optimization of the mobility detection handover performance.
- ⑨ For cross-layer integration of MIP and SIP, some information exchange and coordination is necessary. For example, IP address acquisition at the new location could be complemented at the network layer and is then notified to the SIP. Such an information exchange process could improve the wireless resource usage.

It should be noted that the trade-off between the cross-layer complexity and resulting performance improvement is a crucial point in cross-layer design for mobility management.

Fig. 2.8 Cross-layer handover optimization at the mobile node



2.4 Analysis of Typical Mobility Management Technologies

2.4.1 Data Link Layer: Mobility Management in Cellular Systems

The cellular mobile communication network can be used as a typical representation of the mobility management reference models described above. Mobility management of this network is realized at the link layer.

The architecture of the 2G cellular mobile communication network is divided into the home network and the visiting network, which are composed of three subsystems: the MT, the BSS, and the network subsystem (NSS), as shown in Fig. 2.9.

The BSS consists of the BTS and the BSC. The network is composed of many cells, each with their own BTS. The NSS includes the MSC, the VLR, the HLR, the AuC, and the equipment identity register (EIR). According to the network reference model, the BSS is an access point, and the MSC and the GMSC act as the domain mobility controllers, while the EIR, HLR/AuC, and MSC/VLR form the register servers. The EIR stores the International Mobile Equipment Identifier (IMEI) and its available attributes, while the subscriber authority data are kept in the AuC, and

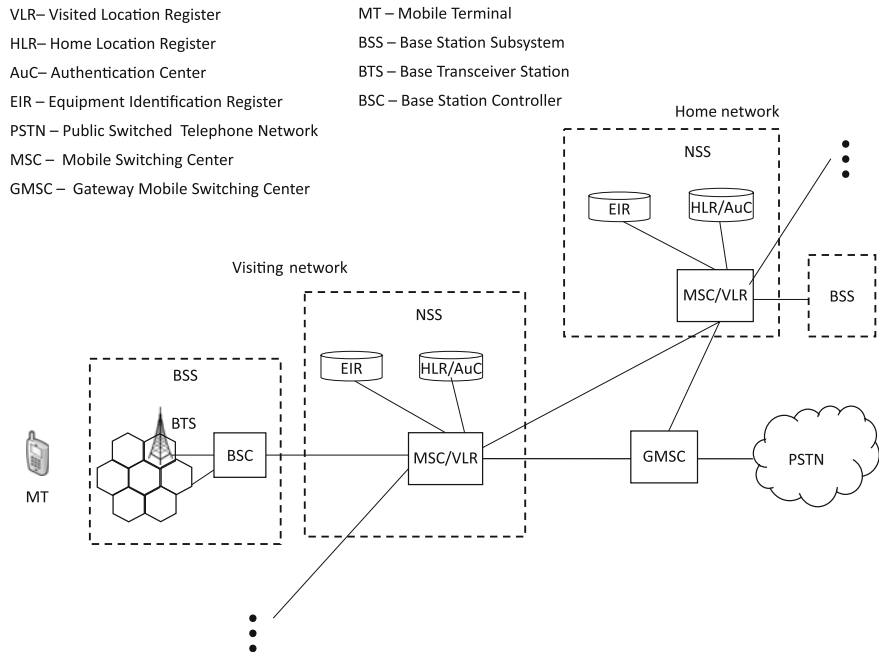


Fig. 2.9 Architecture of 2G cellular mobile communication network

the HLR is used to record the subscriber’s service subscription data and to back up the authority data stored in the AuC. The HLR and VLR form location servers with a two-level database structure. The HLR stores the current subscriber MSC/VLR location information, and the VLR stores the current subscriber LA information. The typical mobile management protocol used is GSM MAP. It is used to exchange messages among MSC, HLR, and VLR and is involved in registration management and location management functions.

According to the mobility management protocol reference model above, from the data plane viewpoint, the cellular mobile network supports mobility primarily in the physical layer (mainly regarding wireless signal measurement) and the link layer. From the control plane viewpoint, mobility management in the cellular network provides the security mechanism, location management, and handover control functions. Table 2.6 summarizes the key points of these control functions. A more detailed description of the mobility management functions in cellular mobile communication networks can be found in Chap. 5.

2.4.2 Network Layer: MIP

MIP, including MIPv4 [26, 27] and MIPv6 [28, 29], supports mobility at the network layer. In the MIP network, an MN, such as a mobile host or a mobile router, can change its network access point, and maintain ongoing communications. Along with the MN, MIPv4 defined two network entities for mobility support: the HA and the FA, as shown in Fig. 2.10.

According to the network reference model, the HA works as a domain mobility controller, a registration server and a location server. The FA is a domain mobility controller. Both the HA and the FA need to maintain related records of MNs, which is similar to the behavior of the HLR and VLR in cellular mobile networks. The

Table 2.6 Mobility management functions in cellular mobile communication networks

Control function	Key points
Security mechanism	• AAA function is completed through signaling interaction between the SIM (subscriber identity module) card and AuC; • The AAA procedure mainly involves the mobile station (SIM), HLR/AuC, and MSC/VLR. The related information of the subscribers is stored at these network elements
Location management	• Two-level database composed of HLR and VLR; • Location information is also stored in the SIM card; • Supports location update and paging
Handover control	• Supports four handover types: intracell handover, intraBSC (RNC) handover, inter-BSC within same MSC/VLR handover and inter-MSC/VLR handover; • Supports hard handover, soft handover, and softer handover

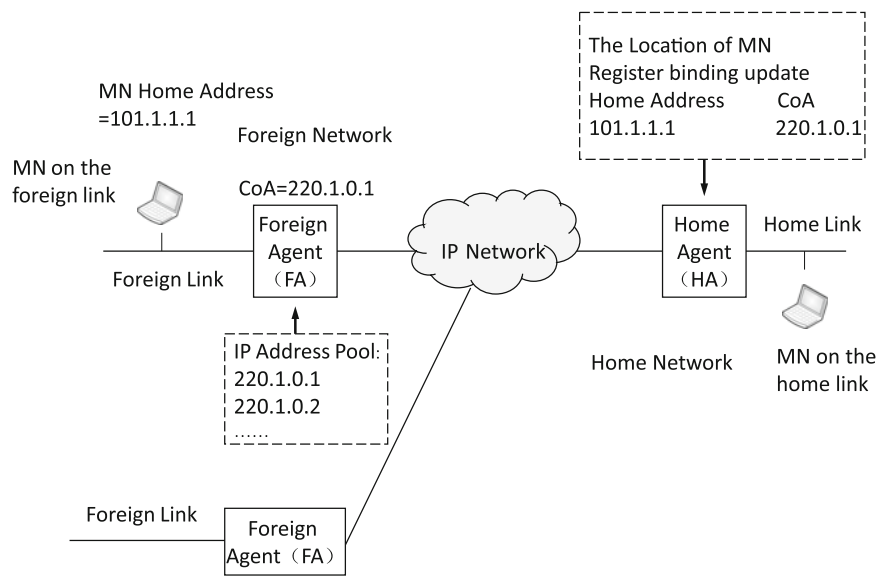


Fig. 2.10 Functional entities in MIPv4

mobility management protocols include agent discovery (an extension based on the Internet Control Message Protocol (ICMP)) and location registration messages.

From the data plane perspective, MIP supports mobility at the network layer. This mobility support is independent of the data link-layer protocols and the physical transmission medium and is transparent to the upper layers and the mobile users.

From the control plane perspective, MIP provides security management, location management, and handover control functions, as shown in Table 2.7. More detailed descriptions of the MIP control functions can be found in Chap. 6.

Table 2.7 Mobility management functions of MIP

Control function	Key points
Security mechanism	• Provides Security Association (SA) between different network entities, e.g., between MN and HA, between MN and FA, and between FA and HA
Location management	• HA maintains binding information of MNs' HoA and CoA and records location information of MNs; • Location registration in MIPv4 and binding update in MIPv6 are used for location update functions; • Paging function is not supported in standard MIP, but may be supported in some extension protocols
Handover control	• Handover occurs when the MN moves across subnets; • Supports fast handover, smooth handover, and seamless handover

2.4.3 Transport Layer: MSCTP

SCTP is a transport-layer protocol defined by the IETF in RFC 2960 [30], which inherits some congestion and traffic control mechanisms from TCP and has some new features, such as multi-homing, multi-streaming, and selective acknowledgment (SACK) to provide a reliable transport service for the upper layers. Mobile SCTP (mSCTP) is an SCTP extension protocol with a dynamic address reconfiguration (DAR) extension [31]. The multi-homing feature, together with the DAR extension, make mSCTP the typical mobility support protocol at the transport layer. Because mobility supported by mSCTP is based on the end-to-end semantics, only the mobility nodes act as mobility entities, while no modifications are needed for any of the network infrastructure devices.

According to the control plane of the protocol reference model, mSCTP has no security mechanism or location management functions and supports only handover control. The key points of the control functions are shown in Table 2.8. More detailed information about mSCTP can be found in Chap. 7.

2.4.4 Application Layer: SIP

Although SIP was initially proposed as an application-layer control protocol for multimedia sessions [32], it has an inherent personal mobility support capability and can be extended to support terminal mobility and service mobility. Therefore, SIP becomes the typical mobility support protocol at the application layer.

Table 2.8 Mobility management functions of mSCTP

Control function	Key points
Security mechanism	• Provides critical data privacy, registration, authentication and signaling integrity to a certain extent, based on TLS/SCTP, SCTP/IPSec, and S-SCTP
Location management	• No location management function and must be integrated with SIP, MIP, DDNS or RSerPool to support location management functions
Handover control	• Multi-homing feature enables the mSCTP endpoint to be configured with multiple IP addresses, one of which will be chosen as the primary address; • DAR extension provides the capability for address addition, address deletion and primary address reconfiguration to be performed dynamically; • The seamless handover procedure can be described as follows: with the movement of the MN, the newly acquired IP address is added as a secondary address, is then set as the primary address used for data transmission, and is deleted when it becomes unavailable

SIP defines the user agent (UA) and various SIP servers (including the proxy server, the redirect server, the location server and the registrar), as shown in Fig. 2.11. The proxy server acts as the domain mobility controller. The functions of the location server and the registrar are obvious. The functions of the mobility management are included in the SIP extensions.

According to the protocol reference model, the SIP is located at the application layer in the data plane. It has security mechanism, location management, and very limited handover control functions from the control plane viewpoint, as listed in Table 2.9. A more detailed description of the SIP control functions can be found in Chap. 8.

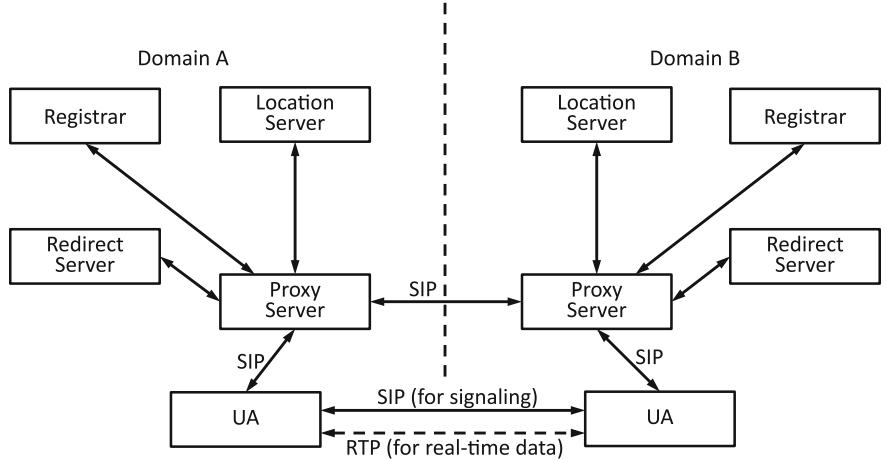


Fig. 2.11 Architecture of the SIP system

Table 2.9 Mobility management functions of SIP

Control function	Key points
Security mechanism	• Registrar is responsible for registration and authentication; • The mobile terminals use the information carried in REGISTER messages for registration and authentication; • HTTP digest authentication method is adopted; • Transport-layer security (TLS) is used for security transmission
Location management	• User location information is updated through location registration; • Location server is responsible for maintaining user-related location information; • Redirect server and proxy server are responsible for finding the current location of the user based on interaction with the location server; • No paging support in the basic protocol, but paging can be realized based on cross-layer integration of SIP and L2 paging;
Handover control	• Supports handover through re-INVITE message; • Handover function is rather limited, only supporting single-side handover based on UDP

References

1. Chen S, Shi Y, Hu B (2007) Mobility management theory and technology. *J Commun* 28 (10):123–133
2. IETF RFC5201 (2008) Host identify protocol
3. IETF RFC5533 (2009) Shim6: level 3 multihoming shim protocol for IPv6
4. IETF RFC4295 (2006) Mobile IPv6 management information base
5. Jain AK, Shanbhag D (2012) Addressing security and privacy risks in mobile applications. *IT Professional* 14(5):28–33. doi: [10.1109/MITP.2012.72](https://doi.org/10.1109/MITP.2012.72)
6. La Polla M, Martinelli F, Sgandurra D (2013) A survey on security for mobile devices. *IEEE Commun Surv Tutor* 15(1):446–471. doi:[10.1109/SURV.2012.013012.00028](https://doi.org/10.1109/SURV.2012.013012.00028)
7. Yihua Zhu (2005) Mobility management in wireless mobile networks. Posts & Telecom Press, Beijing
8. Bar-Noy A, Kessler I (1993) Tracking mobile users in wireless communications networks. In: *Proceedings of IEEE INFOCOM*, pp 1232–1239
9. Pitoura E, Samaras G (2001) Locating objects in mobile computing. *IEEE Trans Knowl Data Eng* 13(4):571–592. doi:[10.1109/69.940733](https://doi.org/10.1109/69.940733)
10. Chakravorty R, Vidales P, Patnapongpibul L, Subramanian K, Pratt I, Crowcroft J (2003) On inter-network handover performance using mobile IPv6. University of Cambridge Computer Laboratory, Technical Report. <http://www.cl.cam.ac.uk/coms/publications.htm>. Accessed 28 March 2014
11. Siddiqui F, Zeadally S (2006) Mobility management across hybrid wireless networks: trends and challenges. *Comput Commun* 29(9):1363–1385
12. IETF RFC3753 (2004) Mobility related terminology
13. Mouly M, Pautet MB, Foreword By-Haug T (1992) *The GSM system for mobile communications*. Telecom Publishing, Washington, DC
14. UMTS (Universal Mobile Telecommunications System) handover. <http://www.umtsworld.com/technology/handover.htm>. Accessed 28 March 2014
15. Xie J, Wang X (2008) A survey of mobility management in hybrid wireless mesh networks. *IEEE Netw* 22(6):34–40. doi:[10.1109/MNET.2008.4694172](https://doi.org/10.1109/MNET.2008.4694172)
16. Chen Y, Kowalik K, Davis M (2009) MeshScan: performance of passive handoff and active handoff. In: *IEEE international conference on wireless communications and signal processing, WCSP 2009*, pp 1–5. doi: [10.1109/WCSP.2009.5371523](https://doi.org/10.1109/WCSP.2009.5371523)
17. Derong Chen, Jiaru Lin (1999) *Digital mobile communication systems*. Beijing University of Posts and Telecommunications Press, Beijing
18. Ekiz N, Salih T, Küçüköner S et al. (2006) An overview of handoff techniques in cellular networks. *Int J Inf Technol* 2(2)
19. Kaur D, Kumar N (2013) Performance analysis of handoff in CDMA cellular system. *Int J Comput Technol* 9(3):1119–1126
20. <http://wireless.cs.tku.edu.tw/~cychang/2G.pdf>. Accessed 28 Mar 2014
21. Wang Q, Abu-Rgheff MA, Akram A (2004) Design and evaluation of an integrated mobile IP and SIP framework for advanced handoff management. *IEEE Int Conf Commun* 7:3921–3925. doi:[10.1109/ICC.2004.1313287](https://doi.org/10.1109/ICC.2004.1313287)
22. Le L, Li G (2007) Cross-layer mobility management based on mobile IP and SIP in IMS. In: *IEEE international conference on wireless communications, Networking and mobile computing, WiCom 2007*, pp 803–806. doi: [10.1109/WICOM.2007.207](https://doi.org/10.1109/WICOM.2007.207)
23. Prior R, Sargento S (2007) SIP and MIPv6: cross-layer mobility. In: *12th IEEE symposium on computers and communications, ISCC 2007*, pp 311–318. doi: [10.1109/ISCC.2007.4381552](https://doi.org/10.1109/ISCC.2007.4381552)
24. IETF draft-sjkoh-msctp-01 (2005) Mobile SCTP (mSCTP) for IP handover support
25. Wang Q, Abu-Rgheff MA (2003) A multi-layer mobility management architecture using cross-layer signalling interactions. In: *5th European personal mobile communications conference*. doi: [10.1049/cp:20030253](https://doi.org/10.1049/cp:20030253)
26. IETF RFC5944 (2010) IP Mobility support for IPv4, Revised

27. IETF RFC3344 (2002) IP Mobility support for IPv4
28. IETF RFC6275 (2011) Mobility support in IPv6
29. IETF RFC3775 (2004) Mobility support in IPv6
30. IETF RFC2960 (2000) Stream control transmission protocol
31. IETF draft-ietf-tsvwg-addip-sctp-08 (2004) Stream control transmission protocol (sctp) dynamic address reconfiguration
32. IETF RFC 3261 (2002) SIP: session initiation protocol

Chapter 3

Mobility Objects and Technologies

The mobility management technology is enriched with various mobility objects, including terminal mobility, personal mobility, service mobility, and network mobility, as introduced in Sect. 1.2.2.

In this chapter, we discuss the various technologies used to support the different mobility objects, as shown in Fig. 3.1.

3.1 Terminal Mobility

Terminal mobility is the most important mobility object and thus has prevalent requirements. As the network environments continue to evolve, the technical principles of mobility management also change accordingly. This section introduces the different terminal mobility support technologies based on an analysis of the technical principles involved.

3.1.1 *Host-Based Mobility Versus Network-Based Mobility*

Depending on whether the mobile nodes (MNs) participate in the mobility management functions, the terminal mobility can be classified into host-based mobility and network-based mobility.

In host-based mobility, it is the MN that is responsible for routing the traffic to its current location through mobility-related signaling exchanged between the MN and the location server when the MN changes its IP address. In network-based mobility, the network is responsible for routing the data traffic to the new location of the MN and does not require the MN to participate in the mobility-related signaling.

Mobile Internet Protocol (MIP) is an example of the host-based mobility management protocol with minimal network involvement. It was in fact the first

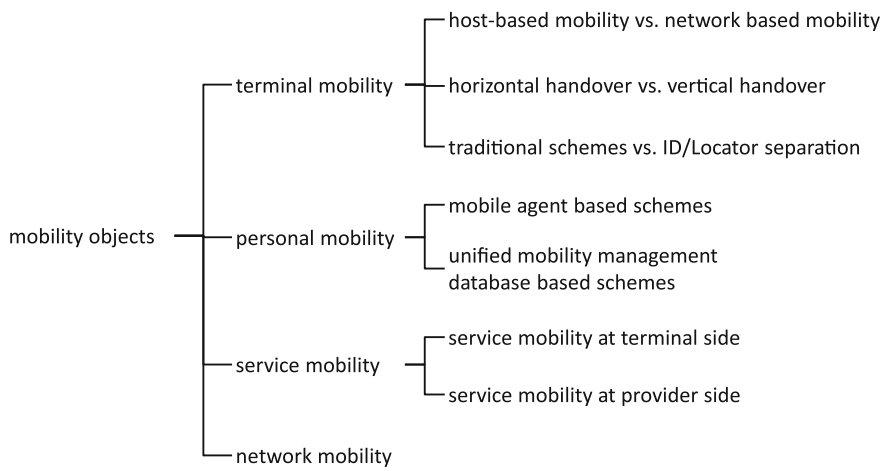


Fig. 3.1 Mobility objects and supporting technologies

mobility management protocol that was proposed for the Internet. The MN participates in the mobility management functions in MIP by updating its location binding information in the home agent (HA) and the correspondent node (CN). The data traffic destined for the MN will then be sent via the tunnel that ends at the MN.

Network-based mobility is proposed based on the considerations of deployment and performance. From the deployment perspective, MIP requires modifications of the MNs’ protocol stack. This is likely to become a major obstacle for deployment of MIP. From the performance perspective, MIP is a macro-mobility management protocol that faces scalability problems when the number of MNs grows. This is because each MN participates in the mobility management signaling procedure and tunneled data transmission, and the signaling messages and tunneled data are exchanged via the wireless links.

Proxy Mobile IPv6 (PMIPv6) is an example of a network-based mobility management protocol. The network takes charge of the host movement detection process, binding updates on behalf of the MNs [1]. The Local Mobility Anchor (LMA) and the Mobile Access Gateway (MAG) are the network entities introduced by PMIPv6 for mobility management. PMIPv6 does not require any modification of the mobile node and thus provides better deployment and mobility performance when compared to host-based mobility protocols. PMIPv6 can be considered to be a derivative of MIPv6 because it uses and extends both Mobile IPv6 signaling and HA functionality. Figure 3.2 shows the differences between the handover processes in MIPv6 and PMIPv6. And Table 3.1 summarizes the comparisons between host-based mobility and network-based mobility.

Table 3.1 shows comparison between host-based mobility management and network-based mobility management.

In PMIPv6, the LMA is responsible for maintaining the MN’s reachability state and acts as a topological anchor point for the MN’s home network prefix(es). The MAG on the new access link, upon detection of the MN on its access link, will

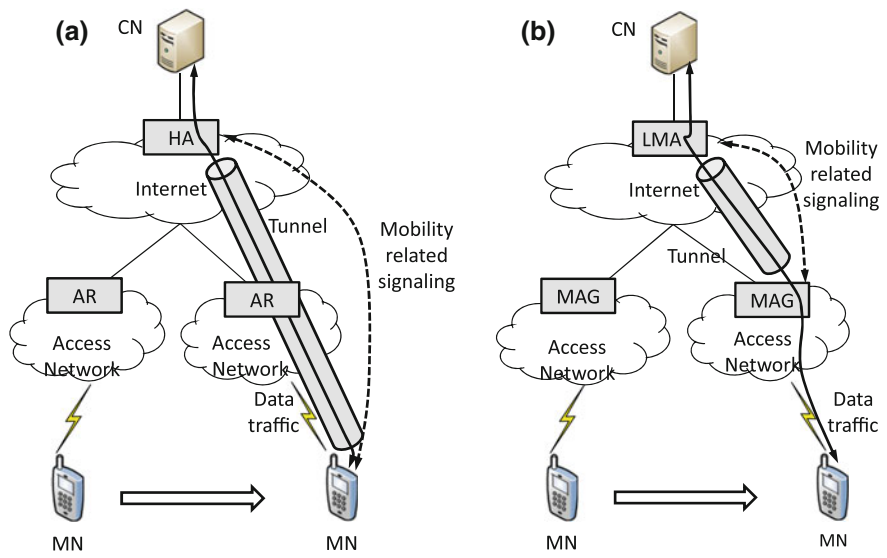


Fig. 3.2 Differences between handover in MIPv6 and in PMIPv6. **a** Handover in MIPv6. **b** Handover in PMIPv6

Table 3.1 Comparison between host-based and network-based mobility management

	Host-based mobility management	Network-based mobility management
Mobility support part	Mobile node and network	Network
Mobility scope	Global mobility	Local mobility
Advantages	Supports global mobility across heterogeneous access technologies	Does not require modification of the MN; Reduced tunneling overhead and mobility-related signaling message exchanges via wireless links [2]; Low handover latency
Disadvantages	Requires modification of the protocol stack of the MN; Large signaling overhead for location updates; Requires tunneling over wireless links	Only supports local mobility
Typical technologies	MIPv4 MIPv6 HIP	PMIPv6 Intra-domain mobility management in cellular systems

signal the LMA to update the binding state. The LMA then creates a binding cache entry and sets up an endpoint for the bidirectional tunnel to the present MAG. Any packet that the MN sends to CN will then be received by the MAG and will be sent to its LMA through the bidirectional tunnel. However, in some cases, traffic sent from a CN that is locally connected to the MAG may not be received by the LMA and may be routed locally by the MAG. There can be multiple LMAs in a single PMIPv6 domain, each serving a different group of MNs.

The MN may be an IPv4-only, IPv6-only, or dual-stack (IPv4/v6) node. PMIPv6 can support mobility for IPv6 nodes without host involvement by extending the MIPv6 [3] signaling messages between a network node and an HA. However, if the MN moves outside a PMIPv6 domain, MIPv6 is then required for mobility management.

3.1.2 *Horizontal Handover Versus Vertical Handover*

Handover is the critical technology that supports session continuity. As described in Sect. 2.3.3, horizontal handover and vertical handover are the handover types depending on whether the access technologies involved are homogeneous or heterogeneous.

A handover between two access points (APs) that are using the same access technology is called horizontal handover, such as the handover between different APs in a wireless local area network (WLAN), and the handover between different base stations (BSs) and mobile switching centers (MSCs) in GSM systems. A vertical handover is performed between APs that use different access technologies, such as a handover between a universal mobile telecommunications system (UMTS) BS and a WLAN AP.

In addition, there are some other differences between vertical handover and horizontal handover. Table 3.2 summarizes these differences.

Vertical handover can be classified further into upward vertical handover and downward vertical handover [4]. The basis for such a classification is the wireless overlap network environment, comprising heterogeneous access networks with different coverage areas. Figure 3.3 portrays an overlap network and shows the different handover types within this network.

The handling of vertical handover is composed of three stages: system discovery, handover decision, and handover execution. System discovery is used to discover the network systems that are currently available. Handover decision determines the handover target network and the handover triggering time. Finally, handover execution implements the handover action.

(1) System discovery

In the system discovery stage, the MN searches and discovers the wireless networks that are currently available. The time cost of the system discovery process

Table 3.2 Differences between horizontal handover and vertical handover

	Horizontal handover	Vertical handover
Access technologies used	Homogeneous	Heterogeneous
Reasons for incurring handover	Physical location change because of movement	Physical location change because of movement; Changes in access technology because of quality of service (QoS) or user preferences
Handover decision criteria	RSS and related indicators; Channel availability	Typical multi-criteria decision problem based on application-related, network-related, and user-related factors
Handover control side	Generally triggered by the network with terminal assistance	Both the network and the user can trigger the handover actively

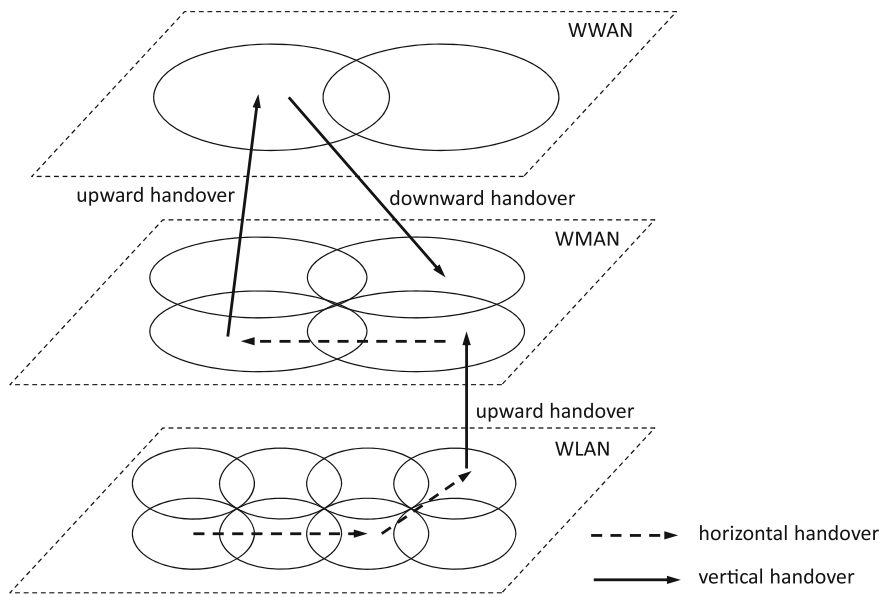


Fig. 3.3 Wireless overlap network with vertical handovers

and the energy cost of the mobile terminal are two important but contradictory parameters in system discovery.

The most commonly used methods for system discovery can be divided into two types:

- System discovery based on messages broadcast over the different network interfaces

Various wireless access networks broadcast messages continuously and periodically, such as the system broadcast message in GSM and the beacon frame in WLANs. A mobile terminal with multiple network interfaces learns about the available access networks through such broadcast messages, which are received at its different network interfaces.

One method simply makes all the interfaces listen continuously to ensure that the shortest possible time is taken for system discovery, but this incurs considerable energy costs at the mobile terminal. For energy-limited terminals, some improvements are necessary.

In the improved method, the mobile terminal opens its network interfaces to search for available networks periodically [5]. The interval used for opening of the interfaces at the terminal should be carefully determined to achieve a trade-off between the system discovery efficiency and the energy cost.

- System discovery based on a location service server (LSS) [6, 7]

The LSS records the information for the various wireless access networks located in different regions. It can therefore provide the surrounding available network systems with their bandwidths, delays, and other parameters based on the geographical location (assisted by the Global Positioning System) of the mobile terminal.

This method can save considerable energy at the terminal. However, the LSS is required to maintain precise and detailed information for every network system, incurring large costs in terms of database establishment and maintenance.

(2) Handover decision

Handover decision determines the handover target network and the handover triggering time. Therefore, it is one of the most important factors affecting the handover performance.

Vertical handover decision is a typical multi-criteria decision problem, where the network availability, status and performance, the current application characteristics and requirements, and the capabilities and preferences of the users/terminals should all be considered together.

Table 3.3 summarizes the network-related, application-related, and user-related factors involved in the vertical handover decision.

Because of the diversity of the factors involved, along with the different units that are used, dedicated algorithms must be designed for the vertical handover decision. The following methods have been proposed:

- Markov decision process based methods [8–13]

The Markov decision process (MDP) is applied to model optimization problems where the decision epochs follow a probability distribution. This is one of the most commonly used mathematical tools for provision of an optimal handover decision policy for vertical handover decisions.

Table 3.3 Factors involved in the vertical handover decision

	Factors involved in vertical handover decision	
	Static factors	Dynamic factors
Network-related	Network operator, network configuration, network type, network coverage area, geographical information, typical bandwidth and delay	Network availability, signal strength, error rate or loss rate, traffic load, available bandwidth and delay
Application-related	Application characteristics, service fees, and security	Resource allocation and priority
User-related	User profile, user preferences, and terminal capabilities	User movement velocity, and historical information of location and user behavior

- A detailed introduction to MDP-based vertical handover decision methods can be found in Sect. 4.4.2.
- **Multi-criteria decision-making-based methods**
The vertical handover decision problem is often formulated as a multi-criteria decision-making (MCDM) problem to select a handover target access network from the candidate networks based on multiple metrics. The weighted sum-based method [14–16], the decision function-based method [17], the analytic hierarchy process (AHP)-based method [18, 19], and the technique for order of preference by similarity to ideal solution (TOPSIS)-based method [20] have all been used in the existing research. A detailed introduction to MCDM-based vertical handover decision methods can be found in Sect. 4.6.
 - **Fuzzy reasoning-based methods [21, 22]**
Fuzzy logic is a theory that deals with uncertainty. The rationale of its application to vertical handover exists in its capability to analyze imprecise data, such as the received signal strength (RSS) and the traffic load of the network. Fuzzy logic can be used directly for vertical handover decisions based on fuzzy inference systems or can be combined with MCDM algorithms, where fuzzy logic is only used to evaluate whether or not the current network is still able to provide the required QoS. A detailed introduction to fuzzy logic-based vertical handover decision methods can be found in Sect. 4.7.1.
 - **Game theory-based methods**
The ability of a mobile terminal to connect simultaneously to multiple wireless access networks introduces new challenges in terms of resource allocation among the mobiles and thus in vertical handover decision making. The vertical handover problem can be seen as a competition between the active participants (the users and the networks), where the users are trying to get the best access network at the lowest cost while the networks are trying to maximize their incomes (on short- and/or long-term scales). The vertical handover decision

problem can then be modeled using various game theory approaches, e.g., cooperative games, non-cooperative games, hierarchical games, and evolutionary games [23].

Examples of the application of game theory to vertical handover decisions can be found in [24–27].

Another problem in vertical handover is the handover control mode. As introduced in Sect. 2.3.3, commonly used handover control modes include network-controlled handover (NCHO), mobile-assisted handover (MAHO), and mobile-controlled handover (MCHO).

Vertical handover could be triggered by the network, or by the terminal, or by both the network and the terminal. The NCHO mode is obviously not suitable for vertical handover, because the performance parameters of the various networks, which are more conveniently acquired via the mobile terminal on the related interfaces than via the network, must be considered in the handover decision. In the MAHO mode, the network controls the handover to improve network resource usage efficiency. For vertical handover across heterogeneous access networks that belong to the same network operator, MAHO is a suitable choice for load balancing and other performance objectives. However, for handover across heterogeneous access networks that belong to different network operators, the MAHO mode is obviously not suitable. When considering the trends for increasingly intelligent terminals, MCHO is another possible choice for user-initiated active handovers, such as active handover from a network with poor performance to a network with better performance for improved QoS.

(3) Handover execution

Handover execution is responsible for the transfer of the ongoing session from the old access point to the target network access point. Because vertical handover is executed across heterogeneous access technologies, the vertical handover protocols run at the network layer and above. MIP, mSCTP (mobile stream control transmission protocol) and SIP (session initiation protocol) are the classical vertical handover protocols at the network layer, the transport layer, and the application layer, respectively. Chapters 6–8 are dedicated to the detailed introduction of these protocols. Table 3.4 compares their vertical handover functions.

It should be noted that the vertical handover protocols described above were originally designed for wired networks. Improvements will be necessary when they are applied to wireless overlay networks, taking the following issues into consideration:

- The high error rates in wireless networks;
- Wireless networks show obvious differences between the network characteristics before and after handover;
- The authentication and registration procedure for users moving across heterogeneous access networks, and the handover delay introduced by this procedure, must also be considered.

Table 3.4 Comparisons between vertical handover functions of MIP, mSCTP, and SIP

Protocol layers	Network layer	Transport layer	Application layer
Protocol	MIPv4/MIPv6	mSCTP	SIP
Network entity requirement	HA (and foreign agent (FA) in MIPv4)	No	Proxy server, registrar and redirect server
Modification of protocol stack of MN	Yes, the MN's network layer must support MIP	Yes, the MN's transport layer must support mSCTP	Yes, the MN's application layer must support SIP
Soft handover support	No	Yes. Soft handover is supported because of the IP diversity support capability based on the multi-homing feature.	Yes. Soft handover is supported based on the JOIN header field extension
Application transparency	Yes	Yes	No
Handover decision function	No built-in handover decision function, relies on the dedicated handover decision mechanism	Built-in handover decision function or dedicated mechanism	No built-in handover decision function, relies on the dedicated handover decision mechanism

Another technology that should be mentioned here is media independent handover (MIH) [28]. MIH has been standardized by the IEEE 802.21 working group to enable the optimization of vertical handovers between heterogeneous IEEE 802 systems and may also ease handovers between IEEE 802 systems and cellular systems [28].

MIH provides a framework that allows higher protocol layers to interact with the lower layers to provide session continuity but without dealing with the specifics of each access technology [29]. MIH defines an access technology independent abstract interface between the upper layers (L3 and above) and the lower layers (L2 and below). This abstract interface can be used to hide the heterogeneity of the underlying access technologies from the upper layers, and thus lead to optimized handover performance.

Figure 3.4 shows the MIH reference model and the related service access points (SAPs).

MIH defines a new entity called the MIH function (MIHF), which is composed of three types of services that enable handovers between heterogeneous access links, including:

- The media independent event service (MIES), which provides event classification, event filtering, and event reporting corresponding to dynamic changes in the link characteristics, link status, and link quality.

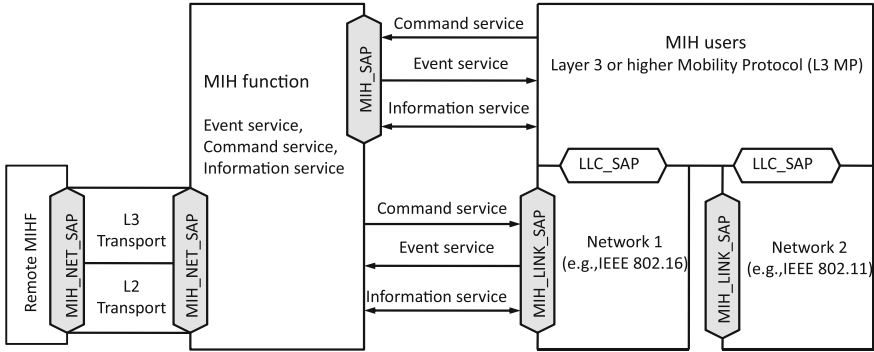


Fig. 3.4 MIH reference model and SAPs. Reprinted from IEEE std 802.21 [28]

- The media independent command service (MICS), which enables MIH users to manage and control the link behavior that is relevant to handovers and mobility.
- The media independent information service (MIIS), which provides details of the characteristics and services provided by the serving and neighboring networks. This information enables effective system access and effective handover decisions.

Some SAPs are also defined in MIH to provide interfaces between the different MIH entities, as shown in Fig. 3.4, and they include [28]:

- **MIH_SAP**: the media independent interface between the MIHF and the MIH user (MIHU) to provide the MIHUs with access to the services of the MIHF.
- **MIH_LINK_SAP**: the technology-dependent interface between the MIHF and the link layer. One SAP is defined for each link layer technology. These SAPs help the MIHF collect the link information and control the link behavior during handovers.
- **MIH_NET_SAP**: a media independent interface that provides transport services for Layer 2 (L2) and Layer 3 (L3) MIH message exchange with remote MIHFs.

MIH can be applied in the wireless overlay network environments for vertical handover optimization. Figure 3.5 shows an example of such an application.

In the scenario shown in Fig. 3.5, an MN with multiple network interfaces is enabled with MIH capabilities. The MN can thus connect to the Internet via different access networks. Assuming that the MN is currently using access network 1, when the link event (e.g., *Link_Going_Down* means that the link conditions are degrading and the connection loss is imminent) from the link layer indicates that the wireless link to access network 1 will become unavailable, the following operations are necessary for a mobile-initiated handover to maintain the session continuity:

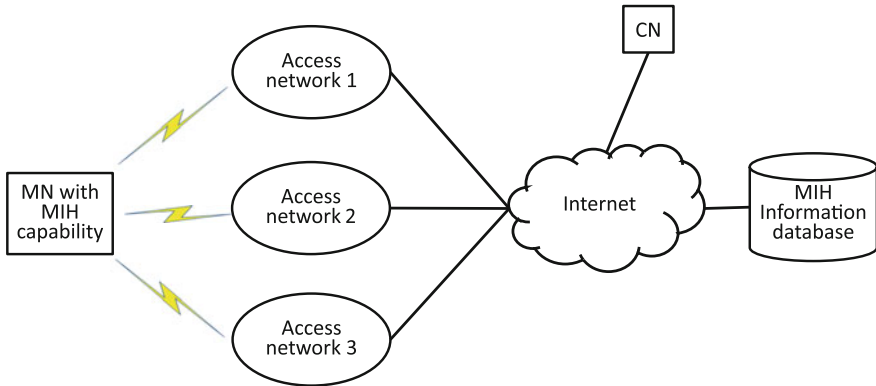


Fig. 3.5 Example of application of MIH in wireless overlay networks

- The MN queries the MIH information server about the available neighboring networks;
- The MN sends a request to the access network to determine the availability of the resources at the candidate networks;
- The MN decides on the target handover network;
- The MN invokes the target network interface through MIH command;
- Specific mobility management protocol procedures are carried out between the MN and the target network to complete the handover.

3.1.3 Traditional Mobility Management Versus ID/Locator Separation

In the current TCP/IP stack, the IP address is overloaded with the semantics of both the host identifiers and the locators [30]. From the application layer viewpoint, the IP address identifies the host, and it is used in the application and transport layers. From the network layer viewpoint, the IP address indicates the current topological location of the host. The dual roles of the IP address make it difficult to support mobility.

The Host Identity Protocol (HIP) [31] supports an architecture that decouples the transport layer (e.g., Transmission Control Protocol (TCP), User Datagram Protocol (UDP)) from the network layer (IPv4 and IPv6) by using public/private key pairs, rather than the IP addresses, as the host identities. One consequence of this type of decoupling is that new solutions for network layer mobility and host multihoming become possible. When a host uses HIP, the overlying protocol sublayers (e.g., the transport layer sockets) are bound to the representations of these host identities, and the IP addresses are actually only used for packet forwarding.

HIP is a host-based ID/locator split scheme, which separates the host identity property from the dual role of the IP addresses through the introduction of a new namespace called the host identity namespace, which is formed by all possible Host Identifiers (HIs). From the perspective of the layered TCP/IP protocol stack, HIP introduces a new layer into the stack.

The locator/ID separation protocol (LISP) [32] is a network layer-based protocol that enables the separation of IP addresses into two new numbering spaces: end-point identifiers (EIDs) and routing locators (RLOCs). No changes are required either to the host protocol stacks or to the “core” of the Internet infrastructure. LISP can be deployed incrementally and offers mobility, multi-homing, and traffic engineering benefits to early adopters.

The architecture of LISP, which is based on ID/locator separation, is convenient for mobility management. The ID is used for session establishment, and the locator information is used for data packet forwarding by looking up the ID to a locator mapping system. When the host moves, only the location information of the host is changed. The network must update the host identity information and the location information of the mapping system to ensure session continuity.

3.2 Personal Mobility

Unlike the terminal mobility above, which considers a single terminal changing its access point, personal mobility focuses on changes in APs when the user is using different terminals.

The two most important aspects of personal mobility support are communication reachability and service individuation.

The critical concept for communication reachability is to treat the user, rather than the terminal, as the communication endpoint. A unique identification should be defined for each user. This identification should then be mapped to the application-specific or terminal-specific identification used in communications.

The most commonly used technology for service individuation is mobile agent technology, which eases the migration of the user profile, the user preferences, and the service profile with the movement of the user, and thus provides consistent and individual service environments.

The existing solutions for personal mobility support can be divided into two types: mobile agent-based schemes and uniform mobility management database-based schemes. These schemes will be introduced in Sects. 3.2.1, and 3.2.2, respectively.

3.2.1 Mobile Agent-Based Schemes

Mobile agent-based personal mobility support is based on software agent technology, which is widely used in the mobile computing field. Depending on whether the agent migrates or not, we have static agents and mobile agents. Mobile agents show autonomy and have active features with migration and cloning operations. Mobile agent technology is often used to provide interoperability between heterogeneous systems.

The references [33–39] describe the classical personal mobility support solutions that are based on mobile agent technology. The agents defined in these solutions can be summarized as:

(1) User agents

User agents can be further classified into static user agents and mobile user agents. The static user agent resides in the user's home network. The mobile user agent resides in the network that the user is currently visiting and is responsible for:

- Following the user's movement and providing services to the user;
- Reporting the user location information to the static user agent in the home network;
- Redirecting communications to a suitable terminal based on terminal availability, user preferences, and application requirements;
- Formatting conversions and adaptations of protocol messages and application content based on the terminal's capability.

(2) Application agents

Application agents are deployed to provide service access during movement. They migrate and are cloned for remote service access when visiting other networks.

(3) User profile agent

The user profile agent is responsible for storage of the user profile information. When the user moves, the user profile agent migrates, following the user to configure the corresponding user profile information in the networks that are visited.

3.2.2 Unified Mobility Management Database-Based Schemes

The unified mobility management (UMM) database was proposed by previous researchers [40–43] for personal mobility support in heterogeneous network environments.

The UMM concept was originated from abstraction of the mobility management entities and functionalities. The abstract mobility management model is defined to capture the common structures of the mobility management activities conducted in different networks [42], but removes the differences in terms of the specific protocols and network entities.

The abstract mobility management model has three major elements: the mobility manager, the originating call control point (O-CCP), and the destination call control point (D-CCP).

The mobility manager manages the MM database, which contains the subscribed service information, and the presence and location information of the users. This information can be accessed virtually from any serving system to provide ubiquitous service support.

The mobility manager also provides mobility management control logic. This logic is normally triggered by a request message from an O-CCP. Certain requests from an O-CCP, such as user authentication, can be fulfilled by the mobility manager alone via database lookup. In many other cases, such as registration and call delivery, the mobility manager satisfies requests by contacting a D-CCP.

The O-CCP and D-CCP can be regarded as the communication source and its destination.

Based on the abstract mobility management model, the UMM-based mobility management architecture is defined. UMM is introduced as a new entity that provides mobility management, security management, and unified user profile management functions for various networks and protocols. UMM is endowed with protocol interaction and translation functions, and thus appears to be simply a native mobility manager for each protocol. For example, UMM provides the functions of a home location register (HLR) for cellular systems, whereas it provides SIP location server functions for SIP-based networks.

The UMM-based mobility management architecture is composed of four elements, including [42]:

- The UMM database: stores the location information of all terminals that belong to each user, and the subscribed service information.
- The core logic server (CLS): provides protocol-independent service logic, such as registration, call delivery, and authentication.
- The protocol-dependent logic server (PDLS) related to a specific protocol: implements protocol-specific service logic.
- Common operations (COPS): provides an interface between the protocol-independent CLS and the protocol-dependent PDLS for mapping and conversion of request messages and response messages.

3.3 Service Mobility

As mentioned in Sect. 1.2.2, service mobility is a type of mobility that is applied for a specific service. This involves the ability of the moving object to use a particular (subscribed) service, irrespective of the user's locations and the terminal used for that purpose [44].

Service mobility support can be divided into two types, depending on where the mobility occurs. The first type focuses on the scenario where the services run in a device and can be transferred to other devices, which we call service mobility at the terminal side. The second type covers scenarios where the service is provided by the device and can be transferred to other devices, which we call service mobility at the provider side.

3.3.1 Service Mobility at the Terminal Side

Service mobility at the terminal side emphasizes service portability, service environment portability, and personal service subscription portability across all the available domains, operators, and terminals.

The virtual home environment (VHE) [45] is an important technology that is used to support this type of service mobility. VHE is a concept for personal service environment (PSE) portability across the network boundaries and between the terminals. The VHE concept is that users are consistently presented with the same personalized features, user interface customization, and services on any network and at any terminal (within the capabilities of the terminal and the network), wherever the user may be located [45].

The PSE describes how the user wishes to manage and interact with the communications services. The PSE is a combination of a list of subscriptions (detailing services provided), the preferences associated with these services, terminal interface preferences, and other information related to the user's experience of the system. Within the PSE, the user can manage multiple subscriptions, e.g., both business and personal, multiple terminal types, and express location and temporal preferences. The PSE is defined in terms of one or more user profiles [45].

One important feature of VHE is the creation of terminal-independent and network-independent services based on a standardized application interface. For this purpose, a series of toolkits can be used for service creation, including Open Service Access (OSA), Customized Applications for Mobile network Enhanced Logic (CAMEL), Mobile Execution Environment (MExE), and the USIM Application Toolkit (USAT).

The critical aspects of implementation of VHE are the standardization of the terminal service execution environments, the distributed storage and management of the user profiles, and media adaptation in mobility.

SIP is another technology used to provide service mobility at the terminal side, which makes it possible for the user to use different terminals through heterogeneous networks or different network operators to access their network services based on their unique user identification. The three critical issues in SIP-based service mobility are:

- The portability of the user identification information and the service profile information;
- Ensuring the QoS of the current session when the user changes his access point to the network;
- Ensuring that the user can access his subscribed services when his access point to the network changes.

To solve the above problems for service mobility support, SIP adopts the user profile portability carrying or network storage methods, the COMET method [46] extension, and the integration of AAA (authorization, authentication and accounting) and the SIP registration.

3.3.2 *Service Mobility at the Provider Side*

Service mobility at the provider side has attracted considerable research effort in recent years, particularly with the development of grid computing and cloud computing technologies.

In this type of service mobility, the definition of service is not limited to traditional network services and can be extended to any of the network resources. The service providers may also be the mobile terminals themselves, the content providers, or virtual machines besides the traditional network service providers. The service mobility may be incurred by user movement, network topology changing, or the network resource scheduling policies, such as load balancing. More precisely, the service mobility here can be defined as the capability to move the service function(s) among the service execution nodes while maintaining an ongoing service [47].

Reference [48] proposes the Service Mobility Virtual Network (SMVN) for seamless virtual machine mobility in the cloud computing context. Reference [47] provides a reference model for service mobility management in the virtual environment and defines a control protocol between the virtualized platform and the managing controller to realize service mobility. This capability is eagerly anticipated for the realization of more scalable and reliable telecommunication networks. The authors of [49] consider each service as an individual network node. The SIP is adopted to build a pervasive grid communication environment and a hierarchical SIP architecture is proposed to reduce the rising handoff frequency caused by the increasing numbers of network nodes. Reference [50] introduces a Service Roaming Protocol (SRP) as a mechanism to manage the service mobility in wireless

grid environments. SRP's main advantage over the existing mobility management algorithm is its context-awareness. The SRP algorithm uses the mobile user's geographical location, the terminal's capabilities, the user's personal profile, and the resources of other grid devices to provide efficient management of the mobile users and services.

3.4 Network Mobility

As described in Sect. 1.2.2, network mobility refers to the ability of the network, where a set of fixed or mobile nodes are networked together, to change, as a group, the point of attachment to the corresponding network upon the movement of the network itself [44, 51].

The mobile nodes in one mobile network access the Internet through a common access point to save on wireless resources. The mobility management is conducted based on the granularity of the whole mobile network, which can reduce mobility management signaling overheads when compared to those for mobility management of each individual mobile node.

3.4.1 Network Mobility Scenarios

Network mobility is one of the most popular mobility objects because mobile networks have become such a prevalent scenario. The following scenarios can be considered as example applications of network mobility:

- Personal area networks (PANs), in which the terminals belonging to and around the same user connect to an external network through a mobile router;
- In a bus, train, or airplane, the mobile terminals of the passengers connect to the WLAN provided by the vehicles themselves, and then connect to the IP backbone network through mobile routers that are deployed on the vehicles;
- Mobile routers can be deployed on cars or airplanes to provide external network connections to the control units and devices (e.g., dynamic systems, navigational systems, brake systems, and entertainment systems) in the vehicle.
- Individual fighting units connected to a military network;

Generally speaking, mobile networks connect to external networks via wireless links. The topology inside the mobile network remains steady when the network is moving. The mobile network may simply be composed of a mobile router and a host connecting to it, but may also be a complex network that includes multiple IP subnets with a nested topology where a mobile network is connected under another mobile network.

3.4.2 *Technical Solutions*

The critical entity for network mobility support is the mobile router (MR). The MR is the access point to the external network for all the mobile nodes in the mobile network. More importantly, the MR completes the mobility management functions on behalf of all nodes in the mobile network, thus preventing all the nodes from performing IP address acquisitions and location updates simultaneously. Therefore, the signaling overhead for mobility management is reduced.

Different solutions have been proposed for network mobility support. Similar to the classification of terminal mobility, network mobility can be divided into host-based network mobility and network-based network mobility, depending on whether the network mobility functions are deployed at the mobile router or at the network.

(1) Host-based network mobility

Network Mobility Basic Support Protocol (NEMO BSP) [52] is a typical protocol for host-based network mobility support.

NEMO BSP is an extension of MIPv6. In the NEMO BSP architecture, the hosts are in a subnet called the mobile network, with one or more mobile routers (MRs) running NEMO BSP that manage the mobility on behalf of all the hosts.

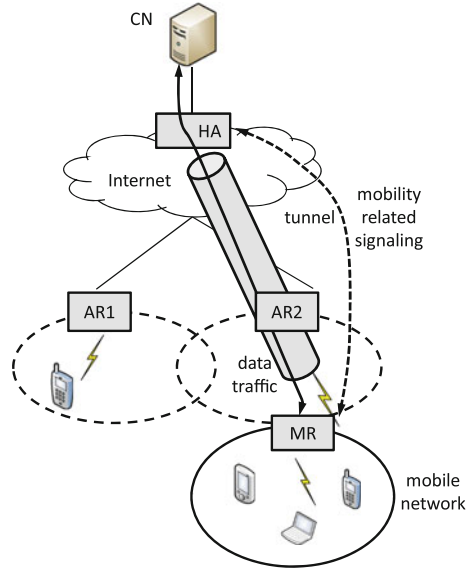
The router can act either as an MN or an MR. When the router acts as an MR, it indicates this to the HA by setting a flag (R) in the binding update message. The MR also includes information about the mobile network prefix in the binding update, so that the HA can forward packets intended for nodes in the mobile network to the MR, which is the default gateway for the mobile network. Communication with the mobile network is conducted through a bidirectional tunnel between the MR and its HA. Thus, the hosts that are connected to the MR are unaware of the mobility.

There are three types of node in a mobile network: a local fixed node (LFN), a local mobile node (LMN), and a visit mobile node (VMN). The solution supports both the mobile nodes and the hosts that do not support mobility in the mobile network.

Mobile networks can also comprise multiple and nested subnets. A mobile network attached to another mobile network forms a nested mobile network. There can be arbitrary levels of nested mobility. The operation of each MR remains the same, no matter whether the MR is attached to another MR or to a fixed AR on the Internet. The nesting level that affects the performance of NEMO BSP can be multiplied when the mobile networks form a hierarchy.

Figure 3.6 illustrates the architecture of the NEMO BSP domain. More details about NEMO BSP can be found in Sect. 6.2.3.

Fig. 3.6 Architectural overview of NEMO BSP



(2) Network-based network mobility

NEMO-enabled PMIPv6 (N-PMIPv6) [53] is an example of a protocol used for network-based network mobility support.

N-PMIPv6 is a network mobility support protocol based on PMIPv6. It fully integrates the mobile networks in PMIPv6-localized-mobility domains [53]. N-PMIPv6 presents two remarkable features in terms of network-based mobility support and improved handover performance.

The main idea of N-PMIPv6 consists of extending the PMIPv6 domain to accommodate mobile networks. Both the original fixed infrastructure (i.e., the MAGs) in PMIPv6 and the mobile networks (i.e., the MRs) belong to the same domain. With N-PMIPv6, an MN attached to a mobile network is also part of the PMIPv6 domain [53].

Figure 3.7 illustrates the architecture of the N-PMIPv6 domain. The mMAG (moving MAG), as shown in the figure, is a new entity defined by N-PMIPv6. It is actually an MR that performs similar functions to those of the MAGs in PMIPv6 while also being mobile.

With regard to the mobility management functions, N-PMIPv6 manages the mobility of the mMAG in the same way that PMIPv6 manages the mobility of the MN. From the viewpoint of an MN attached to the mMAG, this mMAG behaves as a fixed MAG for the N-PMIPv6 domain. To deliver IPv6 packets addressed to an MN attached to the connected mMAG, the LMA functionality is extended to support recursive lookups in its binding cache. The LMA obtains the mMAG to which the MN is attached in the first lookup, and searches for this mMAG in the second lookup to find the associated fixed MAG. Using this information, the LMA

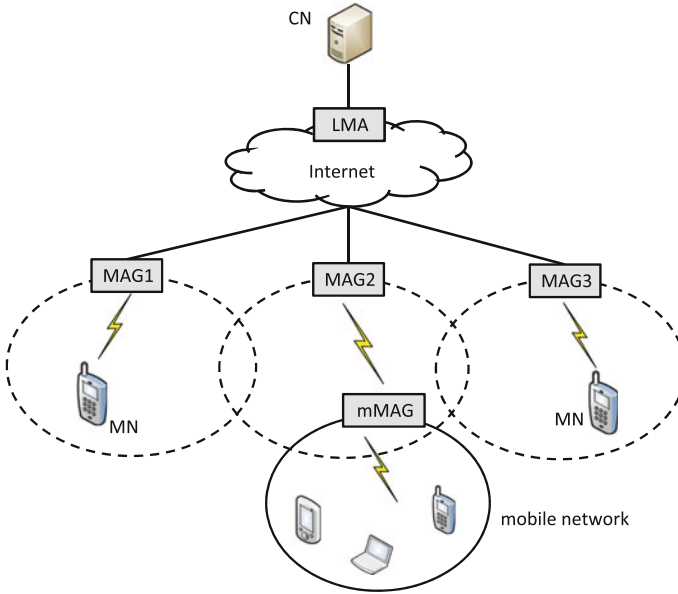


Fig. 3.7 Architectural overview of an N-PMIPv6 domain

can encapsulate the received packets to be sent toward the mMAG through the appropriate fixed MAG. Then, the mMAG can forward the data packets to the MN.

Some other solutions have also been proposed for PMIPv6-based network mobility optimization. A network-based network mobility support scheme (N-NEMO) was proposed in [54]. A tunnel splitting scheme is used in N-NEMO to differentiate the inter-MAG mobility and the intra-MAG mobility. N-NEMO reduces the signaling costs significantly and enhances the efficiency and scalability of the network mobility in the PMIPv6 context. Reference [55] proposed a PMIP-based NEMO protocol with an efficient buffering scheme to prevent packet loss. The efficient buffering scheme manages the buffers with an Automatic Repeat reQuest (ARQ) mechanism.

3.4.3 Routing Optimization in Network Mobility

Routing optimization is an important issue in network mobility. This problem is introduced in NEMO BSP because the packets moving to and from the mobile network are delivered through a bidirectional tunnel between the HA and the MR. Inefficient routing and increased header overheads are incurred because of the encapsulation.

3.4.4 Applications of Network Mobility

Network mobility application in an intelligent transport system (ITS) and satellite networks are demonstrated as two practical cases.

(1) Network mobility in ITS

Network mobility is applied in an ITS for convenient communication of the vehicles. A well-known industrial consortium, the Car2Car Communication Consortium (C2C-CC) [56], has defined the communication architecture in its manifesto document [57], as shown in Fig. 3.8. The architecture comprises the following three domains:

- The in-vehicle domain: This refers to a network that is logically composed of an onboard unit (OBU) and (potentially multiple) application units (AUs). An AU can be an integrated part of a vehicle that is permanently connected to an OBU or can also be a portable device such as a laptop, personal digital assistant (PDA), or game pad that attached to (and detaches

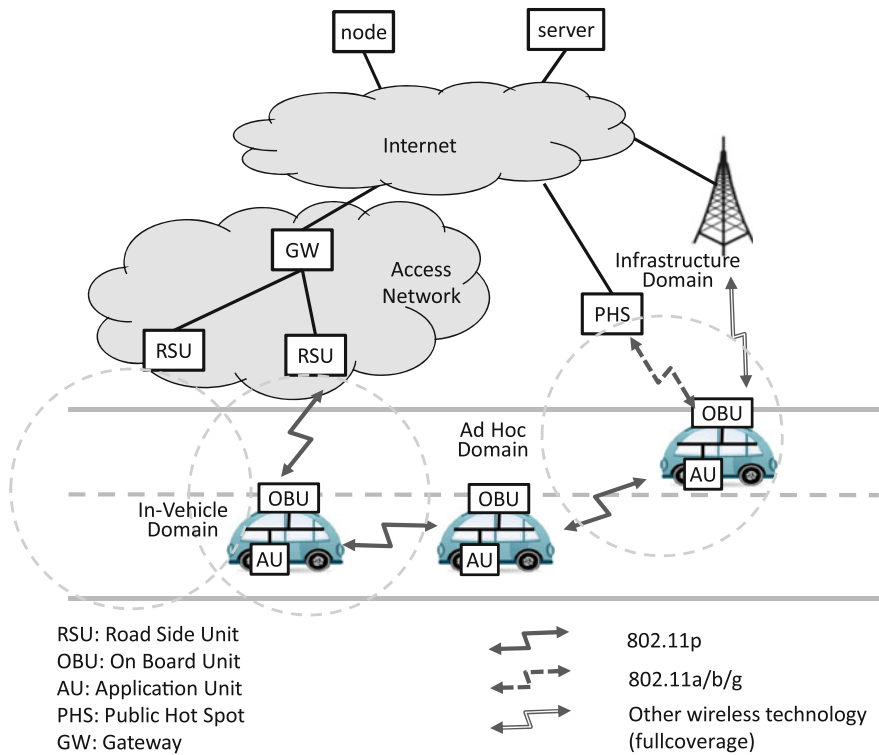


Fig. 3.8 Communication architecture for ITS as defined by C2C-CC. Reprinted from Ref. [57], with the permission from the CAR2CAR Communication Consortium

from) the OBU dynamically. The link between the AU and the OBU can be wired or wireless.

- The ad hoc domain: This domain is composed of vehicles equipped with OBUs and stationary units arranged along the road, known as road-side units (RSUs). The communication between the OBUs or between an OBU and an RSU can be performed in single-hop or multi-hop forms. The wireless links between the nodes in the Ad Hoc domain use the 802.11p standard. Such a scenario is also known as Vehicle-to-Vehicle (V2V) communications.
- The infrastructure domain: This refers to the scenario in which the OBUs are connected to the Internet through public hotspots (PHSs) or other wireless technologies with full coverage, like cellular systems (e.g., GSM, GPRS, UMTS, high-speed downlink packet access (HSDPA), WiMax, 4G). The OBUs are required to be equipped with alternative wireless technologies, such as 802.11a/b/g or cellular radio standards. Such a scenario is also known as Vehicle-to-Infrastructure (V2I) communications.

Based on the communication architecture defined above by C2C-CC, the following network mobility scenarios can be identified:

- The embedded devices in a vehicle, acting as AUs, connect to the OBU to form a mobile network, where the OBU can be deployed as the MR;
- The mobile terminals of the passengers inside a vehicle, acting as AUs, connect to the OBU to form a mobile network, where the OBU can again be deployed as the MR.

For the above mobile networks, and depending on the different communication scenarios, the CN may be:

- Another vehicle, i.e., another MR installed in another vehicle;
- The dedicated nodes that are deployed along the roadside or on the Internet, e.g., the roadside information management nodes or the control center of the service provider over the Internet;
- Any node over the Internet.

In the protocol architecture of the C2C Communication System [57], Mobile IPv6 NEMO is the solution considered by C2C-CC to provide both session continuity and global reachability. As a matter of course, some necessary extensions to MIPv6 NEMO must be proposed to provide better network mobility support because of the complexity of the communication scenarios in ITS, especially when ad hoc communication and nested topologies are involved with NEMO.

(2) Network mobility in satellite networks

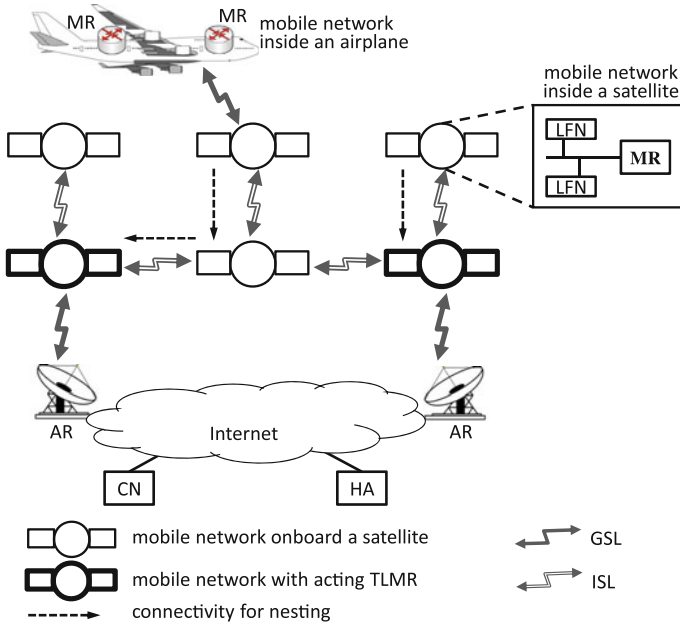


Fig. 3.9 Network mobility in satellite networks. Reproduced from Ref. [58] with the permission from John Wiley & Sons Ltd

A satellite network is a network of satellites and stations on the ground. The satellites connect to each other through inter-satellite links (ISLs) and to the ground stations through ground-to-satellite links (GSLs).

Figure 3.9 shows network mobility in satellite networks, from which the following network mobility scenarios can be derived:

- A mobile network inside an airplane, e.g., the mobile terminals of the passengers connect to the MRs deployed on the airplane to form a mobile network.
- A mobile network inside a satellite, e.g., the onboard IP-enabled devices, acting as local fixed nodes (LFNs), connect to an onboard MR to form a mobile network.
- A nested mobile network, in which one MR connects to another MR through the ISL. Top-level mobile routers (TLMRs) are directly connected to the ARs through the GSLs. The ARs are colocated with the ground stations.

Network mobility support solutions at the different layers have been proposed for the above scenarios in satellite networks [59–65], where the performance optimization and routing optimization schemes are proposed based on the characteristics of the satellite networks.

In particular, for network mobility support of mobile networks inside an airplane, the International Civil Aviation Organization (ICAO) has proposed MIPv6 as

an option for implementation. In addition, network mobility support based on PMIPv6 is also being considered as another option by the ICAO [66].

References

1. IETF draft-sijeon-netext-mmag-pmip-00 (2012) Network mobility support using mobile MAG in proxy mobile IPv6 domain
2. Kong KS, Lee W, Han YH, Shin MK, You H (2008) Mobility management for all-IP mobile networks: mobile IPv6 vs. proxy mobile IPv6. *IEEE Wireless Commun* 15(2):36–45
3. IETF RFC3775 (2004) Mobility support in IPv6
4. Chen WT, Liu JC, Huang HK (2004) An adaptive scheme for vertical handoff in wireless overlay networks. In: *Proceedings of the IEEE Tenth International conference on parallel and distributed systems (ICPADS 2004)*, pp 541–548. doi:[10.1109/ICPADS.2004.1316136](https://doi.org/10.1109/ICPADS.2004.1316136)
5. Stemm M, Katz RH (1998) Vertical handoff in wireless overlay networks. *ACM Mobile Networking (MONET)*, Special Issue on mobile networking in the internet 3(4):335–350. doi:[10.1023/A:1019197320544](https://doi.org/10.1023/A:1019197320544)
6. Inoue M, Mahmud K, Murakami H et al (2003) MIRAI: a solution to seamless access in heterogeneous wireless networks. *IEEE Int Conf Commun (ICC 2003)* 2:1033–1037. doi:[10.1109/ICC.2003.1204509](https://doi.org/10.1109/ICC.2003.1204509)
7. Chen W-T, Shu Y-Y (2005) Active application oriented vertical handoff in next-generation wireless networks. *IEEE Wireless Commun Netw Conf (WCNC 2005)* 3:1383–1388. doi:[10.1109/WCNC.2005.1424718](https://doi.org/10.1109/WCNC.2005.1424718)
8. Sun C, Stevens-Navarro E, Shah-Mansouri V et al (2011) A constrained MDP-based vertical handoff decision algorithm for 4G heterogeneous wireless networks. *Wireless Netw* 17 (4):1063–1081. doi:[10.1007/s11276-011-0335-x](https://doi.org/10.1007/s11276-011-0335-x)
9. Chen Y, Chen H, Xie L et al (2011) A handoff decision algorithm in heterogeneous wireless networks with parallel transmission capability. *IEEE Veh Technol Conf (VTC 2011-Fall)* 1–5. doi:[10.1109/VETECF.2011.6092894](https://doi.org/10.1109/VETECF.2011.6092894)
10. Zhang J, Chan HCB, Leung VCM (2006) WLC14-6: a location-based vertical handoff decision algorithm for heterogeneous mobile networks. *IEEE Glob Telecommun Conf (GLOBECOM2006)* 1–5. doi:[10.1109/GLOCOM.2006.697](https://doi.org/10.1109/GLOCOM.2006.697)
11. Stevens-Navarro E, Wong VWS, Lin Y (2007) A vertical handoff decision algorithm for heterogeneous wireless networks. *IEEE Wireless Commun Netw Conf (WCNC 2007)* 3199–3204. doi:[10.1109/WCNC.2007.590](https://doi.org/10.1109/WCNC.2007.590)
12. Song Q, Jamalipour A (2008) A quality of service negotiation-based vertical handoff decision scheme in heterogeneous wireless systems. *Eur J Oper Res* 191(3):1059–1074
13. Zhu L, Yu FR, Ning B (2010) An optimal handoff decision algorithm for communication-based train control (CBTC) systems. *IEEE Veh Technol Conf (VTC 2010-Fall)* 1–5. doi:[10.1109/VETECF.2010.5594428](https://doi.org/10.1109/VETECF.2010.5594428)
14. Hasswa A, Nasser N, Hossanein H (2005) Generic vertical handoff decision function for heterogeneous wireless. In: *Second IFIP International conference on wireless and optical communications networks (WOCN 2005)*, pp 239–243. doi:[10.1109/WOCN.2005.1436026](https://doi.org/10.1109/WOCN.2005.1436026)
15. Nasser N, Hasswa A, Hassanein H (2006) Handoffs in fourth generation heterogeneous networks. *IEEE Commun Mag* 44(10):96–103. doi:[10.1109/MCOM.2006.1710420](https://doi.org/10.1109/MCOM.2006.1710420)
16. Hasswa A, Nasser N, Hassanein H (2006) Tramcar: a context-aware cross-layer architecture for next generation heterogeneous wireless networks. *IEEE Int Conf Commun* 1:240–245. doi:[10.1109/ICC.2006.254734](https://doi.org/10.1109/ICC.2006.254734)
17. Hong KH, Lee SK, Kim LY et al (2009) Cost-based vertical handover decision algorithm for WWAN/WLAN integrated networks. *EURASIP J Wireless Commun Netw* 2009:15. doi:[10.1155/2009/372185](https://doi.org/10.1155/2009/372185)

18. Balasubramaniam S, Indulska J (2003) Vertical handovers as adaptation methods in pervasive systems. In: The 11th IEEE International conference on networks (ICON2003), pp 705–710. doi:[10.1109/ICON.2003.1266273](https://doi.org/10.1109/ICON.2003.1266273)
19. Song Q, Jamalipour A (2005) A network selection mechanism for next generation networks. IEEE Int Conf Commun (ICC 2005) 2:1418–1422. doi:[10.1109/ICC.2005.1494578](https://doi.org/10.1109/ICC.2005.1494578)
20. Zhang W (2004) Handover decision using fuzzy MADM in heterogeneous networks. IEEE Commun Netw Conf (WCNC 2004) 2:653–658. doi:[10.1109/WCNC.2004.1311263](https://doi.org/10.1109/WCNC.2004.1311263)
21. Guo Q, Zhu J, Xu X (2005) An adaptive multi-criteria vertical handoff decision algorithm for radio heterogeneous network. IEEE Int Conf Commun (ICC 2005) 4:2769–2773. doi:[10.1109/ICC.2005.1494852](https://doi.org/10.1109/ICC.2005.1494852)
22. Zekri M, Jouaber B, Zeghlache D (2010) Context aware vertical handover decision making in heterogeneous wireless networks. In: The 35th IEEE conference on local computer networks (LCN), pp 764–768. doi:[10.1109/LCN.2010.5735809](https://doi.org/10.1109/LCN.2010.5735809)
23. Zekri M, Jouaber B, Zeghlache D (2012) A review on mobility management and vertical handover solutions over heterogeneous wireless networks. Comput Commun 35(17):2055–2068
24. Niyato D, Hossain E (2006) A cooperative game framework for bandwidth allocation in 4G heterogeneous wireless networks. IEEE Int Conf Commun (ICC 2006) 9:4357–4362. doi:[10.1109/ICC.2006.255766](https://doi.org/10.1109/ICC.2006.255766)
25. Niyato D, Hossain E. (2006) Wlc04-5: bandwidth allocation in 4 g heterogeneous wireless access networks: a noncooperative game theoretical approach. IEEE Glob Telecommun Conf (GLOBECOM 2006) 1–5. doi:[10.1109/GLOCOM.2006.637](https://doi.org/10.1109/GLOCOM.2006.637)
26. Haddad M, Altman Z, Elayoubi S E et al (2010) A nash-stackelberg fuzzy q-learning decision approach in heterogeneous cognitive networks. IEEE Glob Telecommun Conf (GLOBECOM 2010) 1–6. doi:[10.1109/GLOCOM.2010.5684318](https://doi.org/10.1109/GLOCOM.2010.5684318)
27. Niyato D, Hossain E (2009) Dynamics of network selection in heterogeneous wireless networks: an evolutionary game approach. IEEE Trans Veh Technol 58(4):2008–2017. doi:[10.1109/TVT.2008.2004588](https://doi.org/10.1109/TVT.2008.2004588)
28. IEEE std 802.21 (2008) Media independent handover services
29. De La Oliva A, Banchs A, Soto I et al (2008) An overview of IEEE 802.21: media-independent handover services. IEEE Wireless Commun 15(4):96–103. doi:[10.1109/MWC.2008.4599227](https://doi.org/10.1109/MWC.2008.4599227)
30. IETF RFC4989 (2007) Report from the IAB workshop on routing and addressing
31. IETF RFC4423 (2006) Host identity protocol (HIP) architecture
32. IETF RFC6830 (2013) The locator/ID separation protocol (LISP)
33. Di Stefano A, Santoro C (2000) NetChaser: agent support for personal mobility. Internet Comput 4(2):74–79. doi:[10.1109/4236.832949](https://doi.org/10.1109/4236.832949)
34. Jung E, Park YJ, Park C (1998) Mobile agent network for supporting personal mobility. In: Proceedings of the twelfth international conference on information networking, (ICOIN-12), pp 131–136. doi:[10.1109/ICOIN.1998.648367](https://doi.org/10.1109/ICOIN.1998.648367)
35. Maniatis P, Roussopoulos M, Swierk E et al (1999) The mobile people architecture. ACM SIGMOBILE Mobile Comput Commun Rev 3(3):36–42. doi:[10.1145/329124.329153](https://doi.org/10.1145/329124.329153)
36. Thanh D et al (1999) Using mobile agent paradigm in mobile communications. In: Ericsson conference software engineering
37. Wang HJ, Raman B, Chuah C et al (2000) ICEBERG: an internet core network architecture for integrated communications. IEEE Pers Commun 7(4):10–19. doi:[10.1109/98.863991](https://doi.org/10.1109/98.863991)
38. Thai B, Wan R, Seneviratne A (2001) Personal communications in integrated personal mobility architecture. In: Ninth IEEE International conference on networks, pp 409–414. doi:[10.1109/ICON.2001.962377](https://doi.org/10.1109/ICON.2001.962377)
39. Thai B, Wan R, Seneviratne A et al (2003) Integrated personal mobility architecture: a complete personal mobility solution. Mobile Netw Appl 8(1):27–36. doi:[10.1023/A:1021115610456](https://doi.org/10.1023/A:1021115610456)

40. Haase O, Murakami K, LaPorta TF (2003) Unified mobility manager: enabling efficient SIP/UMTS mobile network control. *IEEE Wireless Commun* 10(4):66–75. doi:[10.1109/MWC.2003.1224981](https://doi.org/10.1109/MWC.2003.1224981)
41. Isukapalli R, Alexiou T, Murakami K (2002) Global roaming and personal mobility with COPS architecture in SuperDHLR. *Bell Labs Tech J* 7(2):3–18. doi:[10.1002/bltj.10002](https://doi.org/10.1002/bltj.10002)
42. Murakami K, Haase O, Shin JS et al (2004) Mobility management alternatives for migration to mobile internet session-based services. *IEEE J Selected Areas Commun* 22(5):818–833. doi:[10.1109/JSAC.2004.826920](https://doi.org/10.1109/JSAC.2004.826920)
43. Haase O, Xiong M, Murakami K (2004) Multi-protocol profiles to support user mobility across network technologies. In: *Proceedings of the IEEE International Conference on Mobile Data Management*, pp 100–105. doi:[10.1109/MDM.2004.1263047](https://doi.org/10.1109/MDM.2004.1263047)
44. ITU-T Q.1706/Y.2801 (2006) Mobility management requirements for NGN
45. GPP TR22.121 (2002) The virtual home environment (Release 5); v5.3.1
46. IETF draft-ietf-sip-manyfolks-resource-03 (2001) Integration of resource management and SIP
47. IETF draft-yokota-opsawg-virtnw-service-management-02 (2011) Managing service mobility for virtualized networks
48. Asghar NBE, Frikha M, Tabbane S (2011) A virtual network solution for service mobility. In: *IEEE Wireless and Mobile Networking Conference (WMNC), 2011 4th Joint IFIP*, pp 1–6. doi:[10.1109/WMNC.2011.6097233](https://doi.org/10.1109/WMNC.2011.6097233)
49. Wu YS, Liu CF, Chen CN (2010) Service mobility support for pervasive grid based on SIP. In: *2010 IET International conference on frontier computing, theory, technologies and applications*, pp 19–24. doi:[10.1049/cp.2010.0531](https://doi.org/10.1049/cp.2010.0531)
50. Rashid B, El-Dariby M (2007) Towards seamless service mobility for mobile devices communicating within wireless grids. In: *IEEE Canadian conference on electrical and computer engineering (CCECE 2007)*, pp 491–495. doi:[10.1109/CCECE.2007.129](https://doi.org/10.1109/CCECE.2007.129)
51. Chen S, Shi Y, Hu B (2007) Mobility management theory and technology. *J Commun*
52. IETF RFC3963 (2005) Network mobility (NEMO) basic support protocol
53. Soto I, Bernardos C, Calderon M et al (2009) NEMO-enabled localized mobility support for internet access in automotive scenarios. *IEEE Commun Mag* 47(5):152–159. doi:[10.1109/MCOM.2009.4939291](https://doi.org/10.1109/MCOM.2009.4939291)
54. Yan Z (2011) Researches on key technologies for mobility support based on MIPv6/PMIPv6. PhD thesis of Beijing Jiaotong University
55. Lee HB, Min SG, Lee KH et al (2010) PMIPv6-based NEMO protocol with efficient buffering scheme. In: *Proceedings of the 5th International conference on ubiquitous information technologies and applications (CUTE)*, pp 1–6. doi:[10.1109/ICUT.2010.5677637](https://doi.org/10.1109/ICUT.2010.5677637)
56. CAR 2 CAR Communication Consortium (C2C-CC) (2008) <http://www.car-to-car.org/>. Accessed 5 Dec 2013
57. C2C-CC manifesto. <http://www.car-to-car.org/index.php?id=31>. Accessed 2 Dec 2013
58. Shahriar AZM, Atiquzzaman M, Ivancic W (2011) Network mobility in satellite networks: architecture and the protocol. *Int J Commun Syst* 26(2):177–197. doi:[10.1002/dac.1338](https://doi.org/10.1002/dac.1338)
59. Carlos Gonzaga López (2008) GSM on board aircraft. Ari Rantala (TAMK University of Applied Sciences). <http://www.etsi.org/technologies-clusters/technologies/mobile/gsm-onboard-aircraft>. Accessed 2 Dec 2013
60. Chowdhury PK, Atiquzzaman M, Ivancic W (2006) SINEMO: an IP-diversity based approach for network mobility in space. In: *Second IEEE International conference on space mission challenges for information technology (SMC-IT 2006)*, vol 7. p 115. doi:[10.1109/SMC-IT.2006.68](https://doi.org/10.1109/SMC-IT.2006.68)
61. IETF RFC 5522 (2009) Network mobility route optimization requirements for operational use in aeronautics and space exploration mobile networks
62. Bauer C, Zitterbart M (2011) A survey of protocols to support IP mobility in aeronautical communications. *IEEE Commun Surv Tutor* 13(4):642–657. doi:[10.1109/SURV.2011.111510.00016](https://doi.org/10.1109/SURV.2011.111510.00016)

63. Ayaz S, Bauer C, Eddy WM et al (2008) NEMO route optimization solution space analysis and evaluation criteria for aviation. In: IEEE 8th International conference on ITS telecommunications (ITST 2008), pp 139–145. doi:[10.1109/ITST.2008.4740244](https://doi.org/10.1109/ITST.2008.4740244)
64. Qingbo L, Xuejun Z, Liang Z (2009) A mobility management mechanism in aeronautical telecommunication network. In: IET International communication conference on wireless mobile & computing (CCWMC 2009), pp 594–597. doi:[10.1049/cp.2009.2023](https://doi.org/10.1049/cp.2009.2023)
65. Liu Y, Li J, Zhang H (2011) NEMO route optimization for aeronautical passenger communications. In: IEEE 2011 International conference on transportation, mechanical, and electrical engineering (TMEE), pp 2310–2314. doi:[10.1109/TMEE.2011.6199682](https://doi.org/10.1109/TMEE.2011.6199682)
66. ICAO Doc 9896 (2011) Manual for the ATN using IPS standards and protocols; Draft Version 19. http://www.icao.int/safety/acp/ACPWGF/ACP-WG-I-14/IP%208%20DOC9896_rev_19_%20Directory%20Services%20Added.pdf. Accessed 5 Dec 2013

Chapter 4

Application of Mathematical Theories in Mobility Management

In the research on mobility management technology, classical mathematical theories are useful tools for mobility modeling, performance analysis, and algorithm design. This chapter describes the application of these theories in mobility management technology and presents some examples.

4.1 Overview

As discussed in Chap. 2, the critical control functions in mobility management include location management, handover control, interoperability control, and security mechanisms.

Table 4.1 lists the applications of various mathematical theories in the different control functions of mobility management.

4.2 Probability and Statistics Theory

As a mathematical foundation for statistics, probability theory is a branch of mathematics dealing with probability, which is the analysis of random phenomena. It also provides the basis for stochastic processes and queuing theory.

In mobility management research, probability and statistics theory is also a fundamental theoretical tool. Some classical probability distributions are used as important assumptions in mobility analysis. The probabilities of certain events are critical metrics in performance evaluation. In addition, the Bayes' multi-criteria decision making is applied in handover decisions. Detailed introductions to these theories, along with some examples of their applications, will be given in the following sections.

Table 4.1 Applications of mathematical theories in mobility management

Control function	Mathematical theory	Applications
Common assumption	Probability theory	The residence time of a mobile node at a position The call/request arrival rate distribution
	Stochastic processes	Mobility modeling of individuals and groups
Location management	Probability theory	Location management overheads analysis
	Stochastic processes	Use of renewal processes for analysis of residence time and location update time
	Markov processes	Modeling of system behavior for location management performance and overheads evaluation
	Queuing theory	Location operation delay and location database access delay analysis
	Fuzzy logic theory	Location registration efficiency improvement; Location prediction for efficient paging
Handover control	Probability theory	Handover performance evaluation
	Bayes' multi-criteria decision making	Handover decision rules
	Markov processes	Modeling of system behavior for wireless channel allocation modeling during handover
	Markov decision processes	Vertical handover decisions
	Queuing theory	Queuing strategies in channel allocation Handover delay performance analysis
	Multi-criteria decision-making theory	Vertical handover decisions
	Fuzzy logic theory	Handover target network/cell selection Handover triggering time decisions Handover threshold determination

4.2.1 Using Classical Distributions as Assumptions

Some classical probability distributions are used to form important assumptions in the theoretical analysis of overheads and performance in mobility management. For example, the residence time of a mobile node in a particular position is often assumed to follow an exponential distribution, while the calls arriving at mobile nodes are often assumed to follow Poisson distributions.

Of course, the accuracy of these assumptions is important to ensure the accuracy of the corresponding theoretical analysis. The basic classical distributions must be revised in certain cases to depict the statistical properties of the parameters correctly.

For example, incoming calls to a mobile terminal are often assumed to form a Poisson process. Therefore, the interservice time (i.e., the time between the end of

one call and the start of the next call served by the mobile terminal [1]) is often assumed to be exponentially distributed by many researchers.

However, this commonly used assumption is not always correct.

The work of Fang et al. [1] indicated that this assumption is correct only when the busy line effect is negligible. It is possible for a call to arrive while the previous call is still in progress. In this case, the mobile terminal cannot accept the new call (the caller hears a busy tone in this case and may hang up). Thus, the interarrival (or intercall) times for the calls that terminated at the mobile terminal are different from the interservice times. This phenomenon is called the busy line effect. Although the incoming calls do form a Poisson process (i.e., the interarrival times are exponentially distributed), the interservice times may not necessarily be exponentially distributed. Therefore, when the busy line effect is negligible, the interservice time is exponentially distributed. However, when the busy line effect is considered, we should assume that the interservice time is generally distributed [1].

The work of Younas et al. [2] also indicated that this assumption is only applicable in a non-blocking environment such as a telephone network. The assumption does not fit well with the growing traffic on the Internet and the pervasive and mobile cyberspace. In cellular networks, a request blockage occurs when insufficient channels (resources) are available and when the requests tend to arrive in bulk. Therefore, the generalized exponential (GE) distribution [3] is used to model the arrival of requests in this manner [2].

4.2.2 *Defining the Probabilities of Classical Events for Analysis*

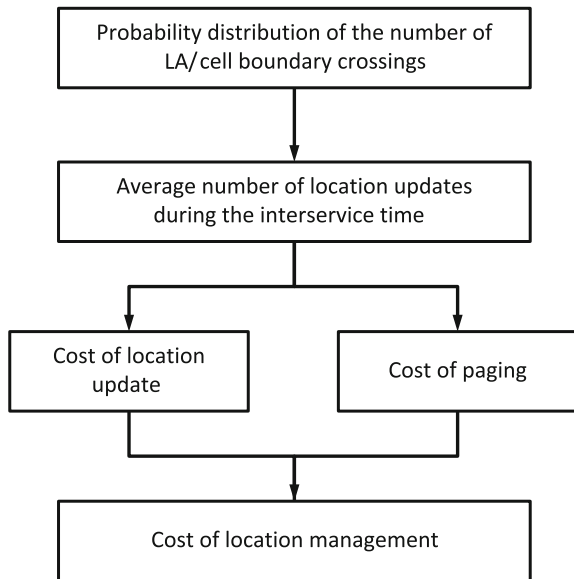
The probabilities of some important events are often critical metrics in mobility management strategy design, performance evaluation, and overhead analysis. For example, in location management, the probability distribution of the number of location area (LA) boundary crossings is extremely important for the design and performance evaluation of the location management strategies. In handover control, the handover success probability, the handover dropping probability, and the new call blocking probability are important metrics in handover performance evaluation.

Here, location management is considered as a detailed example to illustrate the application of probability theory. As described earlier in Sect. 2.3.2, location management strategy research aims to achieve a trade-off between the costs of location updates and those of paging. The total cost of a location management strategy can generally be defined as:

$$C = C_u(N_u, \lambda_u) + C_p(N_p, \lambda_p)$$

where C_u is the cost of the location updates and C_p is the cost of paging. N_u is the average number of location updates that occur under a given location update

Fig. 4.1 Commonly used method for location management cost analysis



strategy, and N_p is the average number of paging messages required under the given paging strategy. λ_u is the signaling cost for each location update operation, and λ_p is the signaling cost for each paging operation.

Figure 4.1 describes the method that is commonly used in location management strategy cost analysis.

As shown in Fig. 4.1, in an analysis of the location management costs, the following parameters should be determined step by step:

- The probability distribution of the number of LA/cell boundary crossings between two consecutive calls that arrive at the same mobile user. This distribution can be derived based on the mobility model assumed for the users and the interservice time distribution.
- The average number of location updates during the interservice time. This number can be derived based on the probability distribution of the number of LA/cell boundary crossings and depends on the different location update strategies that can be used, such as the basic policy, the pointer extending policy, the anchor-based policy, and the circle searching policy mentioned in Sect. 2.3.2.
- The cost of location updates. This is a function of the average number of location updates during the interservice time and the cost of each location update operation.
- The cost of paging. This is related to the location database structure, the location update strategy, and the paging strategy. The average number of location updates may also be included as part of the paging cost analysis.

4.2.3 Use of the Bayes' Multi-criteria Decision Making for Handover Evaluation

The Bayes' multi-criteria decision making is one of the most widely used methods in signal detection technology, which uses statistical decision theory as its mathematical basis.

The Bayes' multi-criteria decision making can be used in handover decisions for definition and estimation of the handover rules [4, 5].

Reference [4] applied the Bayes' multi-criteria decision making to the definition of the handover decision rules for intercell handovers in mobile radio communications. The use of the Bayes' multi-criteria decision making was proposed in preference to the original situation, where the handover decision was made based on direct comparison of the decision criteria to pre-defined thresholds. The Bayes' multi-criteria decision making is applied here in an additional procedure, in which the handover decision is made on the basis of achieving an optimal trade-off between two conflicting criteria: the number of unnecessary handovers and the number of lost calls.

Reference [5] also applied the Bayes' multi-criteria decision making for intercell handovers in mobile radio communications. By considering the handover decision algorithms that compare the differences between the received signal strengths from different base stations to a hysteresis value, Ref. [5] provided a contribution toward optimal estimation of the hysteresis value using the Bayes' multi-criteria decision making.

4.3 Stochastic Processes

A stochastic process is a collection of random variables [6]. This process is often used to represent the evolution of certain random values, or systems, over time.

In mobility management research, stochastic processes are useful tools in descriptions of incoming call arrival and user mobility behavior, and they are also used to model the system states and state transitions in handover, location management, and resource allocation to provide optimized solutions.

It should be noted that only the general applications of stochastic processes are introduced in this section, while subsequent dedicated sections (Sects. 4.4 and 4.5) introduce the application of some specific stochastic processes. Section 4.4 introduces the application of Markov processes to mobility management, and the Markov chain is also often used as an analysis tool in queuing theory, which is introduced in Sect. 4.5.

Detailed introductions to the general applications of stochastic processes can be found in the following sections.

4.3.1 Modeling of Incoming Call Arrival

In mobility management research, the Poisson process is often used to model the arrival of incoming calls and forms an important assumption for the theoretical analysis.

A Poisson process is a stochastic process that counts the number of events and the time that these events occur in a given time interval. The time between each pair of consecutive events has an exponential distribution with parameter λ , and each of these interarrival times is assumed to be independent of other interarrival times [7].

When the incoming call arrivals (or the requests for a particular document on a Web server) are assumed to be a Poisson process, then certain probabilities or features can be derived accordingly [8–12], including the following:

- the interarrival times, which are exponentially distributed;
- the length of time that the mobile terminal stays in a cell (called the cell residence time), which has a general non-lattice distribution;
- the moving distance distribution between the last location update and the next incoming call;
- the probability that boundary crossings occur between two call arrivals.

These probabilities can then be used in the theoretical analysis of the different location update, paging, and handover solutions.

4.3.2 Mobility Modeling and Prediction

Mobility models represent the movements of the mobile users and how their locations, velocities, and accelerations change over time [13].

Mobility modeling plays a number of important roles in mobility management research. It forms an important basis for performance evaluation, overhead analysis, and the simulation and design of location update, paging, and handover processes.

In addition, mobility management solutions have experienced the development from passive solutions to active solutions. Traditional mobility management mechanisms passively follow the movements of the mobile users or terminals and maintain their connections to the network. In active mobility management, mobility predictions can be made based on the most appropriate mobility models. The future location of each mobile user is predicted based on their current location and the laws with regard to their changes in velocity or location. Then, the location updates, paging, and handover strategies can be designed accordingly, based on the different rules such as distance or time.

In most mobility models, the movement of the user or terminal is assumed to be random. Therefore, the stochastic process is a most useful tool for mobility modeling.

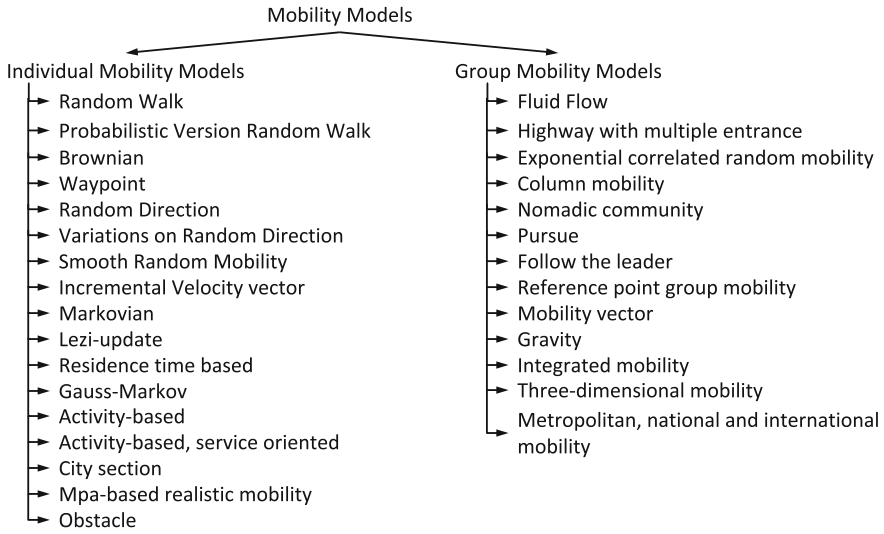


Fig. 4.2 Classical mobility models

Depending on whether the movements of the different users are independent or dependent, the mobility models can be classified into individual mobility models (also known as entity mobility models) and group mobility models [14]. Figure 4.2 shows some classical mobility models.

The most commonly used mobility models are introduced in the following section.

A. Commonly used individual mobility models

In the individual mobility models that are introduced in this section, the movements of the mobile nodes are considered to be completely independent of each other.

Random walk mobility model (including its many derivatives): A simple mobility model based on random directions and speeds.

In this mobility model, a mobile node (MN) moves from its current location to a new location by randomly choosing its direction and speed of travel. The new speed and direction are both chosen from pre-defined ranges, which are $[speedmin, speedmax]$ and $[0, 2\pi]$, respectively. Each movement in the random walk mobility model occurs in either a constant time interval t or over a constant distance traveled d ; at the end of each time or distance interval, a new direction and speed are calculated. If an MN that moves according to this model reaches a simulation boundary, it simply “bounces” off the simulation border at an angle that is determined by the incoming direction. The MN then continues along this new path [14].

Many derivatives of the random walk mobility model have been developed, including the 1-D, 2-D, 3-D, and d-D (d-dimensional) walks. The 2-D random walk

mobility model is of particular interest, because the Earth's surface is modeled with the use of a 2-D representation.

The random walk mobility model is a memoryless mobility pattern, because it retains no knowledge with regard to its past location and speed information [15]. The current speed and direction of any MN are independent of its past speed and direction [16].

The random walk mobility model is a widely used mobility model (e.g., [17–20]) and is sometimes referred to as Brownian motion.

Random waypoint mobility model: A model that includes pause times between the changes in direction and/or speed [21].

In the random waypoint model, the MN initially remains in a single location for a certain period of time (i.e., a pause time). When this time expires, the MN then chooses a random destination in the simulation area and a speed that is uniformly distributed in the $[\text{minspeed}, \text{maxspeed}]$ range. The MN then travels toward the newly chosen destination at the selected speed. Upon arrival at the destination, the MN pauses for a specified time period before starting the process again [14].

The random waypoint mobility model is also widely used (e.g., [22–25]).

Brownian motion mobility model

Reference [3] used Brownian motion with a drift process to model the user mobility under study. A 1-D Brownian motion with a drift process starting at position $X(t_0) = x_0$ at time $t = t_0$ can be described by the following distribution [26, 27]:

$$p_x(x|x_0, t) = \frac{1}{\sqrt{2\pi D(t - t_0)}} \exp \left\{ -\frac{[x - x_0 - v(t - t_0)]^2}{2D(t - t_0)} \right\}, \quad t \geq t_0$$

where the mean and the variance of the process are linear functions of the time difference $(t - t_0)$, i.e.,

$$\begin{aligned} E[x(t)] &= \mu(t) = x_0 + v(t - t_0) \\ \text{Var}[x(t)] &= \sigma^2(t) = D(t - t_0) \end{aligned}$$

Here, D is the diffusion parameter (also called the variance parameter) with units of $(\text{length}^2/\text{time})$. D represents the average location uncertainty [28] of the motion, and the drift velocity v ($\text{length}/\text{time}$) represents the average velocity of a moving user.

Gauss–Markov mobility model: A model that uses a single tuning parameter to vary the degree of randomness in the mobility pattern.

In the Gauss–Markov mobility model, each MN is initially assigned a current speed and direction. At fixed time intervals denoted by n , movement occurs by updating the speed and direction of each MN. Specifically, the values of the speed and direction of the MN at the n th instance are calculated based upon the values of its speed and direction at the $(n - 1)$ st instance and a random variable. At each time interval, the next location is calculated based on the current location, the speed, and

the direction of movement. Also, to ensure that an MN does not remain near the edge of the grid for a long period of time, the MNs are forced away from any edge when they move within a certain distance of that edge [14].

The Gauss–Markov mobility model can eliminate the sudden stops and sharp turns that are encountered in the random walk mobility model by allowing the previous velocities (and directions) to influence the future velocities (and directions) [14].

The freeway mobility model and Manhattan mobility model described in the following are the most commonly used mobility models in vehicular networks.

Freeway mobility model: A model that is often used to model the mobility behavior of vehicles running on freeways.

In the freeway mobility model, each MN is restricted to its lane on the freeway. The velocity of the MN is temporally dependent on its previous velocity. If two MNs in the same freeway lane are within the safety distance (SD), the velocity of the following node cannot exceed the velocity of the leading node [29].

The freeway mobility model is often applied in research on ad hoc networks and vehicular networks.

Manhattan mobility model: A model that is often used to model the mobility of vehicles running in a grid road topology.

The Manhattan mobility model was mainly proposed for movement in urban areas, where the streets are located in an organized manner. In this mobility model, the MNs move in horizontal or vertical directions on an urban map. At an intersection of a horizontal and a vertical street, the MN can turn left, turn right, or go straight ahead with certain probabilities [29, 30].

The Manhattan mobility model is also often applied in research on vehicular networks.

B. Commonly used group mobility models

In the group mobility models that are introduced in this section, the movement of the MNs has cooperative features when they move together. For example, a group of soldiers in a military scenario may be assigned the task of searching a particular plot of land to destroy land mines, capture enemy attackers, or simply work together in a cooperative manner to accomplish a common goal [14].

Fluid flow mobility model: A mobility model that is used to describe the macro-mobility behavior of mobile users.

In the fluid flow mobility model, the movement of each MN follows a random distribution, in which the average speed is v and the movement direction of the MN is uniformly distributed in the $(0, 2\pi)$ range [31].

This model is often used to model the movements of large numbers of mobile users in cellular systems for mobility-related performance analysis. The average boundary crossing times are calculated first.

When the boundary crossing times are calculated, the overall mobility of the mobile users is conceptualized as a fluid flow. The mobile users are assumed to be uniformly distributed across the whole system with a density of ρ . If the length of

the cell boundary is L , then the cell crossing times of a mobile user in a particular time unit could be defined as follows:

$$C = \rho v L / \pi$$

This is then used to derive the average location update rate for location management, the handover times for handover control, and the channel allocation number for resource management.

Reference point group mobility (RPGM) model: A group mobility model where the group movements are based upon the path traveled by a logical center [14].

In the RPGM model, each group has a logical center (a group leader) that determines the group's motion behavior, including their overall direction and speed. Each node within the group has a speed and a direction that is derived by randomly deviating from that of the group leader [16].

The RPGM model is often used in situations where some of the MNs move together and network with each other in an ad hoc manner. One example of this type of mobility is the way that a number of soldiers may move together in a group or platoon. Another example is found during disaster relief operations, where various rescue crews (e.g., firemen, policemen, and medical personnel) form different groups and work cooperatively [32].

The RPGM model is the most classical group mobility models. Other commonly used group mobility models, including the column mobility model, the nomadic community mobility model, and the pursue mobility model, can be implemented as special cases of the RPGM model [14].

4.3.3 Location Management

In location management, the renewal process can be used for the analysis of cell/LA/PA residence times and location update times, which form the basis for location management signaling overhead analysis.

A renewal process is a generalization of the Poisson process, where the holding times take on a more general distribution, while the independent and identically distributed (i.i.d.) properties of the holding times are retained [33].

Reference [34] used the equilibrium renewal process to analyze the cell residence time and the location update time. The cell residence time is defined as the time for which a mobile terminal remains in a cell before it moves into a neighboring cell. The location update time is defined as the time taken for a mobile terminal to cross d cell boundaries and update its location in a movement-based location management scheme (MBLMS), or the time threshold in a time-based location management scheme (TBLMS).

A renewal process is defined by a sequence of independent random variables $T_1, T_2, T_3, \dots$, where $T_2, T_3, \dots$ are a sequence of i.i.d. random variables with a common probability density function (pdf), but T_1 may have a different pdf [34].

The most interesting property described in [34] is the number of renewals in a random period of time t , which is denoted by $X(t)$.

In the modeling of dynamic location management schemes, the T_i values can be cell residence times, or paging area residence times, or location update times. When the T_i values are cell residence times, each renewal represents a cell boundary crossing, and $X(t)$ is the number of cell boundary crossings in a time interval of length t . When the T_i values are paging area residence times, then each renewal represents a PA boundary crossing, and $X(t)$ is the number of PA boundary crossings in a time interval of length t . Finally, when the T_i values are location update times, each renewal then represents a location update, and $X(t)$ is the number of location updates in a time interval of length t [34].

4.4 Markov Process

A Markov process is a stochastic process in which the probabilities of future states are determined only by the present state and not by how the process arrived at the present state. Mathematically, a stochastic process $x(t)$ is a Markov process if for every n and $t_1 < t_2 < \dots < t_n$, the following equation holds:

$$\begin{aligned} P(x(t_n) \leq x_n | x(t_{n-1}), \dots, x(t_1)) \\ = P(x(t_n) \leq x_n | x(t_{n-1})). \end{aligned}$$

Markov processes can be used to characterize a variety of real-world systems. The most typical property of a Markov process is memoryless. This means that one can make predictions for the future of the process based solely on its present state just as well as one could when knowing the full history of the process; i.e., it is only “conditional on the present state of the system, and its future and past are independent” [35, 36].

In mobility management research, a Markov process is a powerful mathematical tool. It is widely used to model systems for performance evaluation, to make vertical handover decisions, and to predict a user’s future location based on their movement history. Detailed introductions to these applications can be found in the following sections.

4.4.1 Performance Evaluation

A Markov process is often used to model the behavior of a system, a network, or a protocol for performance evaluation purposes, which then lays the groundwork for technical improvements. In general, some states are initially defined according to the research issues and a Markov chain is subsequently constructed to describe the

state transition relationship. Then, the steady-state probability of these states can be calculated for the performance evaluation.

In existing research, Markov processes are applied to location management performance and overhead analysis and to wireless channel allocation modeling during handover.

A. Location management performance and overhead evaluation

In location management, the location registration overhead and the paging overhead rely greatly on the border crossing rate. In this case, the “border” may be the cell border, the LA border, the registration area (RA) border, or the paging area (PA) border, depending on the different problems under study and the location management strategies used.

A Markov chain is often used to model the mobility behavior of mobile terminals to derive the border crossing rate for performance and overhead analysis. When setting up the Markov chain model, the states can be defined as the possible locations or the possible movement directions of the mobile terminals. The state transitions between the different states describe the mobility events that may require location update and paging operations. After the steady-state probabilities are obtained, the location management overhead can then be analyzed accordingly.

For example, Ref. [37] used a Markov process in user mobility behavior modeling. An embedded Markov chain is used to model the mobile station registration activities. The state i (where $i > 0$) is defined as the number of RAs, i , that the mobile station has passed. The state transitions between the different states represent the mobile station movement rate to a neighboring RA in the same access network, or the mobile station movement rate to another RA outside the access network. Then, the equilibrium probability of being in state k of the process is obtained for location management overhead analysis.

In another example, Ref. [38] set up a discrete time Markov chain model to capture the mobility and call arrival patterns of the terminal. This was applied in a distance-based location update scheme to determine the location update threshold distance. The state of the Markov chain i (where $i > 0$) is defined as the distance between the current location of the mobile terminal and that of its center cell. The probabilities with which the distance of the terminal from its center cell will increase or decrease are represented by the transition probabilities. Based on the steady state, probabilities are obtained for the location update and paging overhead analysis. An iterative algorithm is then used to determine the optimum location update threshold distance that produces the minimum cost.

More examples can be found in [39, 40].

B. Wireless channel allocation modeling during handover

Channel allocation is a necessary step in the handover procedure, with the aim of achieving a trade-off between the channel utilization and the user’s quality of service (QoS) in terms of blocking probability and drop probability.

In wireless channel allocation modeling using a Markov process, the state space is often defined to represent the different channel occupation states. The state transitions describe the user arrival, user departure, and handover processes based on the different channel allocation strategies. Then, the performance parameters such as the blocking probability, the dropping probability, and the channel utilization can be calculated.

A three-dimensional Markov chain model is used in [41] to analyze the proposed channel preemption mechanism for wireless local area network (WLAN)–cellular integrated heterogeneous networks.

The state is defined by three parameters: the number of MNs outside the WLAN coverage area and employing cellular channels, the number of MNs within the WLAN coverage area and using cellular channels, and the number of MNs within the WLAN coverage area and connecting through WLAN access points (APs). Accordingly, the state transitions into a state define the MN arrival process, while the state transitions out of a state define the MN departure process.

Then, the average number of WLAN users, the average channel utilization, the channel blocking probability, the preemption probability, and the preempted probability are all calculated for the performance evaluation.

A general model using a two-dimensional Markov chain is built in [42] for wireless channel allocation and handover schemes. Based on the examples of Markov models for complete channel sharing and partial channel sharing schemes, a general analysis model is then derived.

For more examples, please refer to [43–45].

4.4.2 Vertical Handover Decision

The Markov decision process (MDP) is one of the most commonly used decision methods in vertical handover decisions. An MDP is a discrete time stochastic control process. It is the optimization decision process for stochastic dynamic systems based on Markov process theory.

An MDP is often used in stochastic order decision problems. The MDP is composed of a set of states, a set of actions, a state transition matrix, and a reward function, as shown in Fig. 4.3.

Consider the process to be in state S_t at time step t . The decision maker may choose any action a_t that is available in state S_t . The process responds at the next time step $t + 1$ by moving into a new state S_{t+1} with the transition probability $P_{at}(S_t, S_{t+1})$ and gives the decision maker the corresponding reward $R_{at}(S_t, S_{t+1})$.

The MDP is applied in vertical handover decision problems, in which the necessary elements of the MDP model, including the decision epochs, states, actions, transition probabilities, and reward functions, are defined as shown in Table 4.2.

Based on the above MDP elements, the vertical handover decision problem can be described as follows. At each decision epoch, the mobile terminal must choose an action based on the current system state. Using this state and action, the system

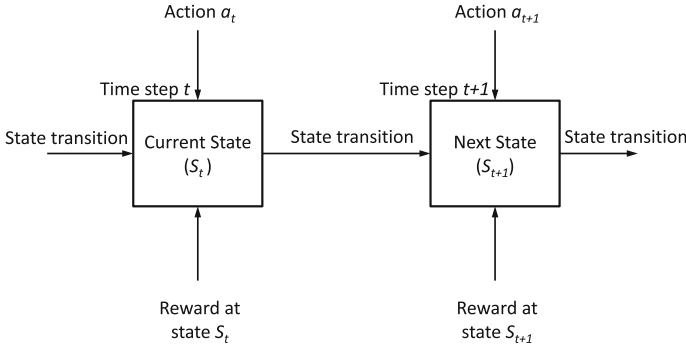


Fig. 4.3 Markov decision process

Table 4.2 Definitions of MDP elements in vertical handover decisions

MDP element	Definition in vertical handover decisions
Decision epoch	The time at which the mobile terminal makes the handover decision
State	The currently available networks with their parameters that are involved in the handover decision (e.g., the QoS that can be provided by each candidate network, and the velocity and location of the mobile terminal)
Action	To select one of the available networks The mobile terminal remains in the current network or handovers to another network, often denoted by the available access network IDs
Transition probability	Given the current state s_t and the chosen action a_t , this is the probability of the next state being s_{t+1}
Reward function	Given the current state s_t and the chosen action a_t , this gives the rewards received by the mobile terminal Often composed of a benefit function and a penalty function Depends on the optimum principles of the vertical handover decision algorithms, e.g., providing the best QoS for the user, or maximizing network revenue without violating the QoS constraints

then evolves into a new state based on a transition probability function. This new state lasts for a period of time until the next decision epoch arrives, and then, the mobile terminal makes another new decision. For any action that the mobile terminal chooses at each state, there is an associated reward [46].

The optimality objective is often defined as obtaining the optimal policy to maximize the total expected reward (with some constraints when a constraint MDP model is used) between the first decision epoch and the termination of the connection, based on the initial state. Dynamic programming algorithms and other optimality theories can be used to find the optimal solutions.

For example, MDPs are used in [47–49] for vertical handover decision problem modeling. When the optimality objective is defined with constraints, the constrained MDP (CMDP) model is applied. The goal of each mobile terminal in [46] is to maximize the expected total reward that it can obtain during the connection,

subject to a constraint on its expected total access cost. Therefore, a CMDP model is used. Besides, semi-Markov decision processes (SMDPs) are also used in [50, 51].

4.5 Queuing Theory

Queuing theory is generally considered to be a branch of operations research and is the mathematical study of waiting lines or queues [52]. In queuing theory, a model is constructed so that the queue lengths and the waiting times can be predicted [52].

Figure 4.4 illustrates the queuing system model. The three fundamental elements in a queuing system are as follows: the server numbers, the arrival rate of the customers, and the serving rate of the servers.

Different queuing models are defined based on the different statistical features of the three elements above, the serving disciplines, and the queuing disciplines. Analysis based on the queuing model is expected to provide parameters including the queue length, the customer waiting times, and the server busy time, which could be the distribution, the statistical parameters, and the parameter estimation, respectively.

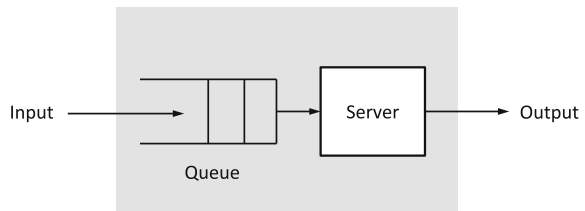
Queuing theory has been widely used in communications and computer areas, e.g., in the design and evaluation of telephone switches, in satellite communication systems, and in operating system scheduling.

In mobility management research in particular, queuing theory can be applied in location management for location operation performance evaluation, and in handover control for resource allocation, access selection, and admission control algorithm design and analysis.

4.5.1 Queuing Strategies in Channel Allocation

Channel allocation is a critical issue in wireless communications. The limited numbers of channels serve both handover access and initial access. The objective of a channel allocation scheme is to achieve high channel utilization efficiency with better trade-offs between the forced termination of handover calls and blocking of

Fig. 4.4 Queuing system model



initial calls. Queuing theory is often used as a mathematical modeling tool for research on queuing strategies for channel allocation, in which the calls that require channels are modeled as the customers and the channels are modeled as the servers. Different queuing models are applied according to different channel assignment principles and queuing strategies.

For example, Ref. [43] discussed a channel assignment scheme for handover calls and initial calls. Based on the author's observation that giving priority to handover calls would not yield a better call completion probability in general, a modified first-in first-out (MFIFO) scheme is proposed. Both handover and initial calls are placed in the same queue. The next free channel is assigned to a handover call or an initial call, based on a pre-defined threshold that determines the best proportion of handover to initial calls in service. Queuing theory is used in [43] to model the proposed channel allocation scheme with performance analysis of the call completion probability.

Reference [53] considered a call admission control (CAC) scheme in cellular multiservice time division duplex (TDD) networks that support both symmetrical traffic (e.g., traditional voice services) and asymmetrical traffic (e.g., Internet data services) with soft handover and handover queuing. The scheme assumes that there is a handover queue provided in each cell with a limited size, which is used to accommodate handover calls that cannot obtain sufficient free channels upon arrival. Based on queuing theory, an analytical model is derived based on the continuous-time multi-dimensional birth-death process and presents performance parameters including the new call blocking probability, the handover failure probability, and the forced termination probability.

Unlike the assumption in most research that a single queue is considered for a whole cell, Ref. [54] proposed a new model with a dedicated queue for each transceiver in the cells of mobile networks with densely deployed microcells and picocells. Each transceiver has a handover queue with a number of channels for handover calls, while the total number of channels is shared equally among the transceivers. Using queuing theory, the new call blocking probability and the forced termination probability of non-blocked calls are derived.

Reference [2] extended the application of queuing theory to channel allocation for context-aware transactions in pervasive and mobile cyberspace. A mobility management scheme was proposed for transactions during mobility to improve the efficiency and throughput of the context-aware transactions. Similar to the channel allocation analysis above, the transaction requests are classified into handover transaction requests (HTR) and new transaction requests (NTR). Unlike the assumption in traditional research that the traffic arrival follows a Poisson process, a GE distribution is used to model the arrival of these requests while considering the bursty nature of cellular traffic. Then, the different queuing modeling systems can be integrated for different requests:

- GE/GE/C/N/FCFS: This queuing system is used to model a dedicated single-class queue that manages only the HTRs, according to the first-come first-served (FCFS) discipline.

- **GE/GE/C/N/HoL:** This queuing system is used to model a cell with HTR and NTR interarrival and interservice times at each channel with GE distributions. The free channels are assigned to the requests in the queue according to the head-of-the-line (HoL) scheduling discipline.
- **GE/GE/C/N/PR with pushout priority queue:** This queuing system is used to handle background traffic (e.g., e-mails, Web browsing) and serves both HTR and NTR. The arriving requests are served under the preemptive resume (PR) service discipline. The arrival of high-priority transactions will push out the background traffic in cases where traffic congestion occurs.

4.5.2 Delay Performance Analysis

In performance analysis for mobility management, the delay is one of the most important measures in both location management and handover control.

Location management relies on the location operations (e.g., location update and paging) based on the location database. Access to the location database becomes one of the key issues that determine the location management efficiency and performance. Queuing theory is used to analyze the database access and operation delays.

In [55], the user is modeled as an $M/M/1/1$ queuing system with only two states: busy or not-busy. The call arrivals that find the user to be not-busy trigger paging operations. Then, the average paging and update costs are analyzed accordingly.

Reference [56] used an $M/D/1$ queuing system with a finite queue length to model the paging process, based on the assumption that the paging messages follow a Poisson arrival process. Then, the average paging delay is derived based on the analysis method of the $M/D/1$ system.

For handover control, the handover delay is one of the most important performance measures. Queuing theory is used as the analytical model to evaluate the delay components at each node.

For example, in [57], the overall handover delay is divided into several components according to the subprocedures used during handover, such as the movement detection delay, the address allocation delay, and the session reconfiguration delay. For each component, the delay incurred is analyzed based on the end-to-end transmission delay of the handover-related signaling messages in this subprocedure. In this step, an $M/M/1$ queuing model is applied to derive the total delay at the wireless interface and the total delay at the wired link.

In [58, 59], queuing theory is applied in analysis of the handover delay for Session Initiation Protocol (SIP)-based mobility management. Because a mid-call handover based on SIP requires only a re-INVITE message that travels from the mobile host (MH) to the destination server through some intermediate network entities, such as base station and some call session control function (CSCF) servers in IP multimedia subsystems (IMSs), major delays occur at the MH, the

intermediate entities, and the destination server because of the queuing of the SIP messages.

The M/M/1 queuing model is assumed for the MH and the intermediate entities, while a priority-based M/G/1 model is assumed for the destination server. The rationale behind these assumptions is that while the MH and the intermediate entities perform dedicated jobs, the destination server may be busy with a variety of jobs other than serving the SIP messages and thus may have a general service time distribution [58]. The handover delay can then be derived based on the waiting time formulas of classical queuing theory.

4.6 Multi-criteria Decision Making

Multi-criteria decision making (MCDM) is a subdiscipline of operations research that explicitly considers multiple criteria in decision-making environments [60].

In mobility management research, MCDM is used in some selection problems that must take multiple factors into consideration to make a synthetic decision, e.g., vertical handover decision, radio resource allocation, access selection, and access control algorithms.

Handover target network selection in a vertical handover decision is a classical application of MCDM problems. As stated in Sect. 3.1.2, vertical handover is the terminology that is used relative to horizontal handover. In traditional horizontal handover decision algorithms, the received signal strength (RSS) is the most frequently used indicator in access selection. However, when considering a vertical handover decision between heterogeneous access technologies, the signal strengths of two heterogeneous access networks are absolutely not comparable. It turns to be a typical MCDM problem, combining the application-related, network-related, and user-related metrics.

Depending on the different decision methods used, the applications of MCDM can be classified into weighted sum-based solutions, decision function-based solutions, analytic hierarchy process (AHP)-based solutions, and technique for order of preference by similarity to ideal solution (TOPSIS)-based solutions.

4.6.1 Simple Weighted Sum-Based Solutions

The weighted sum is the simplest MCDM method. Each optional access network is evaluated with a score, which is given by a weighted sum of all the metrics considered.

Each metric is first assigned a weight value, according to its importance in the selection algorithm. Normalization is then necessary to ensure that the sum of the values in their different units is meaningful. Then, the normalized contribution of each metric is multiplied by its weight. The score of each candidate network is then

calculated as the sum of all the weighted products of the metrics considered. The handover target network, which is the one with the maximum score, is then selected by this process.

For example, Refs. [61, 62] evaluated the candidate network performance based on the weighted sum of several QoS-related metrics such as the cost of service, network security, power consumption, the available bandwidth, and the network connection state.

4.6.2 Decision Function-Based Solutions

Decision function-based solutions are designed based on the concept of the cost function (also defined as the loss function, the utility function, or the indirect utility function in some references). An objective function is defined by mapping an optimization problem with multiple variables onto a real number that intuitively represents some “cost” (or loss, or utility) associated with the problem. A solution with minimized cost (or minimized losses, or maximized utility) is the expected decision.

In vertical handover target network selection, the decision function is often defined based on the cost and the utility that may be introduced by the handover process. The cost defines the negative factors that are expected to be reduced by the handover, such as power consumption, monetary costs, end-to-end delays, and jitter. The utility defines the positive factors that are expected to be increased by the handover, including the data rate, network security, and the handover success rate.

For example, a new vertical handover decision algorithm based on a cost function for wireless wide area network (WWAN)/WLAN integrated networks was proposed in [63]. The cost function considers the data throughput, the power consumption, and the monetary cost. When an MN moves from the WWAN-only area to an area that is covered by both the WLAN and the WWAN, a vertical handover is performed only if the cost of the WLAN is less than that of the WWAN.

4.6.3 AHP-Based Solutions

AHP is a structured MCDM method for organization and analysis of complex decisions. AHP provides a comprehensive and rational framework of the decision problem, to represent and quantify its elements, to relate those elements to the overall goals, and to evaluate any alternative solutions [64].

Briefly, the AHP method is a three-step process [65]:

- Step 1 It decomposes the decision problem into different levels of the hierarchy (identification of the decision criteria).

- Step 2 It compares each factor with all other factors within the same level through a pairwise comparison matrix (e.g., comparing the objectives at the first level and the networks with respect to each objective at the second level).
- Step 3 It calculates the sum of products of the weights obtained from the different levels and selects the solution with the highest sum.

We take the network selection mechanism proposed in [66] as an example. AHP is integrated with the gray relational analysis (GRA) method to realize a trade-off between the user preferences, the service application, and the network conditions. AHP is used to determine the weights of the QoS parameters based on the user preferences and the service application, while the GRA method is used to rank the candidate networks.

Another example is the vertical handover decision method proposed in [67]. AHP is used to decompose the vertical handover target network selection problem into multiple objectives, including maximization of the user device preference, maximization of the application bandwidth, minimization of the jitter, delay and loss, and minimization of the bandwidth fluctuations.

4.6.4 TOPSIS-Based Solutions

TOPSIS is a MCDM method based on the concept that the chosen alternative should have the shortest geometrical distance from the positive ideal solution and the longest geometric distance from the negative ideal solution [68].

For handover target network selection, the logic of using the TOPSIS method is that the chosen network is the network that is closest to the ideal solution and the farthest from the worst case solution [69].

TOPSIS is a compensatory aggregation method that is carried out via the following steps: identification of the weights for each metric involved, normalization of the scores for each metric, calculation of the geometric distance between each alternative network and the ideal alternative, and ranking of the alternatives on this basis to determine the final selection [68].

For example, Ref. [70] integrated TOPSIS with fuzzy logic for network selection. Fuzzy logic is only used to represent the imprecise information of some of the metrics involved (e.g., the user preference). TOPSIS is one method that is then used to determine the ranking order of the available networks.

4.7 Fuzzy Theory

Fuzzy logic theory deals with reasoning that is approximate rather than fixed and exact [71]. Because of its advantages, which consist of allowing imprecise/contradictory inputs, using linguistic variables, permitting fuzzy thresholds, reconciling conflicting

objectives, and increased robustness, it has been widely applied in many fields, ranging from control theory to artificial intelligence.

In mobility management research, fuzzy theory is applied to some issues that require dynamic and adaptive decisions based on multiple factors. For example, a fuzzy control system is often used in vertical handover decisions when considering multiple factors, fuzzy inference can be used to adjust handover thresholds and determine handover times, and fuzzy prediction is used for future RSS and user location prediction, which can be used to improve the handover performance.

Detailed introductions to these applications of fuzzy theory can be found in the following sections.

4.7.1 Handover Decision

In handover control, fuzzy logic can be used for handover decisions, including handover target network/cell selection, handover triggering time decisions, and handover threshold adjustment.

It should be noted that fuzzy logic is often combined with the MCDM concept that was introduced in Sect. 4.6 for advanced algorithms. Fuzzy logic is used in these algorithms to handle the imprecise metrics (e.g., the RSS, the network traffic load) that are involved in the handover decision, and also to combine the MCDM methods simultaneously.

A. Handover target network/cell selection

Numerous handover target network/cell selection algorithms based on fuzzy logic have been proposed, the majority of which have been designed for vertical handovers. Fuzzy logic can be used directly to make handover decisions or can be combined with other MCDM methods.

For example, reference [72] proposed an adaptive multi-criteria vertical handover decision algorithm for the WLAN-UMTS (universal mobile telecommunications system) integrated heterogeneous network environment, with the aim of improving the handover decision accuracy and reducing handover packet loss. The handover decision in this algorithm is made directly by the fuzzy logic system. The network bandwidth, the velocity of the mobile server (MS), and the number of users after handover that was predicted using a modified Elman neural network are used as the input parameters for the fuzzy inference engine after fuzzification. The handover decision result is output according to the fuzzy rule base.

Reference [73] provided an example of the combination of fuzzy logic with MCDM methods for handover decisions. In this algorithm, fuzzy logic is integrated with the AHP, one of the MCDM methods, for alternative vertical handover decisions. Fuzzy logic is only used in the first step to evaluate whether or not the current network is still able to provide the required QoS. This evaluation is made

using the QoS information, the bandwidth, the load, the delay, the jitter, and the bit error rate (BER) as the input parameters. If the current network is no longer able to provide the required QoS for the applications, then the alternative vertical handover will be triggered. AHP is used in the second step to perform the network selection.

B. Handover triggering time decision

In [74], fuzzy logic is used in the horizontal handover triggering decision. The proposed scheme is composed of a gray prediction function and a fuzzy decision system. The gray predictor is used to predict the RSS indication (RSSI) value for the inputs of the fuzzy decision system. Subsequently, the fuzzy decision system calculates the handoff factor (HF) and compares the HF with the handoff factor threshold to make the decision.

Fuzzy logic is used in [75] for microcellular handovers, which have more stringent handover delays because of the more frequent handovers introduced by the smaller cell size.

Two handover decision algorithms based on fuzzy logic were proposed in [75]. The first algorithm uses an adaptive fuzzy predictor to predict the future signal strength based on the signal history or, more precisely, to predict the next incoming signal samples from the BSs under consideration. The handover decision is then made according to the predicted signal strength. The second algorithm uses fuzzy averaging to track the signal trends based on the past and uses this information to make the handover decision.

Reference [76] used fuzzy logic to design a self-adaptive handover decision algorithm for mobile Worldwide Interoperability for Microwave Access (WiMAX) to overcome the ping-pong handover and handover delay issues. Two real-time value criteria of the RSSI and the MS velocity are considered as the fuzzy input parameters. The fuzzy inference system's output parameters consist of three handover parameters: the handover factor, the handover threshold, and the handover margin.

C. Handover threshold determination

The handover threshold defines the conditions that trigger handover operations.

Reference [77] proposed a fuzzy inference system for hard handover thresholds. The system uses the signal strength received at the mobile station and the distance between the mobile station and the base station as inputs, and the output is the defined value for the handover decision.

Fuzzy inference is used in [78, 79] to adjust the soft handover threshold dynamically. The aims of this approach are to increase the thresholds at high traffic loads in order to release the traffic channel for supporting more carried traffic, and to decrease the thresholds at low traffic loads in order to give high traffic channel [78]. The fuzzy inference system in the algorithm outputs a soft handover threshold that depends on two inputs: the number of remaining traffic channels and the number of base stations connected to each mobile station.

4.7.2 Location Management

In location management, fuzzy logic can be used for both location registration and location paging.

In [80–82], a series of fuzzy logic-based location registration schemes were proposed to improve location management efficiency, e.g., to reduce the location management cost, or to make the location management adaptive to the changeable movement patterns of mobile terminals.

In [80], fuzzy logic is applied in distance-based location management schemes to determine a suitable distance threshold. A fuzzy logic controller is used to dynamically adjust the distance threshold for each mobile terminal, so that optimal or near-optimal cost performance can be obtained, regardless of the mobile terminals' changing mobility and traffic patterns.

In [81, 82], fuzzy logic is applied in forwarding pointer-based location management schemes. A dynamic location registration strategy with a fuzzy logic controller is proposed to make the location management scheme adaptive to the changeable movement patterns of the mobile terminals. The number of boundary crossings of a mobile host, the chain lengths of the forwarding pointers, and the location management cost are used as the inputs of the fuzzy logic controller. The chain length variation, which is used to adjust the chain lengths of the forwarding pointers, is given as its output.

The Intelligent Fuzzy Online Location Management Strategy (IFOLMS) was proposed in [83]. IFOLMS consists of two main parts: user movement prediction and intelligent user paging/location based on predicted user locations. The fuzzy location estimator in IFOLMS is used to predict the most possible location of each user at any particular time based on the user's movement history. These predictions are then used to make future paging decisions and to reduce the location management signaling costs.

References

1. Fang Y (2003) Movement-based mobility management and trade off analysis for wireless mobile networks. *IEEE Trans Comput* 52(6):791–803. doi:[10.1109/TC.2003.1204834](https://doi.org/10.1109/TC.2003.1204834)
2. Younas M, Awan I (2013) Mobility management scheme for context-aware transactions in pervasive and mobile cyberspace. *IEEE Trans Ind Electron* 60(3):1108–1115. doi:[10.1109/TIE.2012.2198032](https://doi.org/10.1109/TIE.2012.2198032)
3. Lei Z, Saraydar CU, Mandayam NB (1999) Mobility parameter estimation for the optimization of personal paging areas in PCS/cellular mobile networks. In: 2nd IEEE workshop on signal processing advances in wireless communications, SPAWC'99, pp 308–312. doi:[10.1109/SPAWC.1999.783080](https://doi.org/10.1109/SPAWC.1999.783080)
4. Senadji B, Tabbane S, Boashash B (1994) A handover decision procedure based on the minimization of bayes criterion. In: IEEE 44th vehicular technology conference, pp 77–81. doi:[10.1109/VETEC.1994.345161](https://doi.org/10.1109/VETEC.1994.345161)

5. Senadji B, Boashash B (1997) Estimation of the hysteresis value for handover decision algorithms using Bayes criterion. In: Proceedings of 1997 IEEE international conference on information, communications and signal processing, ICICS, vol 3, pp 1771–1775. doi:[10.1109/ICICS.1997.652300](https://doi.org/10.1109/ICICS.1997.652300)
6. Ross SM (1983) Stochastic processes. Wiley, New York
7. http://en.wikipedia.org/wiki/Poisson_process. Accessed 24 Jan 2014
8. Li TY, Lam KY (2002) An optimal location update and searching algorithm for tracking mobile agent. In: Proceedings of the first international joint conference on autonomous agents and multiagent systems: part 2. ACM, pp 639–646. doi:[10.1145/544862.544892](https://doi.org/10.1145/544862.544892)
9. Zhu Y, Leung VCM (2006) Derivation of moving distance distribution to enhance sequential paging in distance-based mobility management for PCS networks. IEEE Trans Wireless Commun 5(11):3029–3033. doi:[10.1109/TWC.2006.05123](https://doi.org/10.1109/TWC.2006.05123)
10. Lei Z, Rose C (1998) Wireless subscriber mobility management using adaptive individual location areas for pcs systems. Conference record of 1998 IEEE international conference on communications, ICC 98, vol 3, pp 1390–1394. doi:[10.1109/ICC.1998.683054](https://doi.org/10.1109/ICC.1998.683054)
11. Li J, Kameda H, Li K (2000) Optimal dynamic mobility management for PCS networks. IEEE/ACM Trans Netw (TON) 8(3):319–327. doi:[10.1109/90.851978](https://doi.org/10.1109/90.851978)
12. Lin YB (1997) Reducing location update cost in a PCS network. IEEE/ACM Trans Netw (TON) 5(1):25–33. doi:[10.1109/90.554719](https://doi.org/10.1109/90.554719)
13. http://en.wikipedia.org/wiki/Mobility_model. Accessed 24 Jan 2014
14. Camp T, Boleng J, Davies V (2002) A survey of mobility models for ad hoc network research. Wireless Commun Mobile Comput 2(5):483–502. doi:[10.1002/wcm.72](https://doi.org/10.1002/wcm.72)
15. Liang B, Haas ZJ (1999) Predictive distance-based mobility management for PCS networks. In: Proceedings of the joint conference of the IEEE computer and communication societies (INFOCOM), vol 3, pp 1377–1384. doi:[10.1109/INFCOM.1999.752157](https://doi.org/10.1109/INFCOM.1999.752157)
16. Hong X, Gerla M, Pei G et al (1999) A group mobility model for ad hoc wireless networks. In: Proceedings of the ACM international workshop on modeling and simulation of wireless and mobile systems (MSWiM), pp 53–60. doi:[10.1145/313237.313248](https://doi.org/10.1145/313237.313248)
17. Bar-Noy A, Kessler I, Sidi M (1994) Mobile users: to update or not to update? In: Proceedings of the joint conference of the IEEE computer and communications societies (INFOCOM), pp 570–576. doi:[10.1109/INFCOM.1994.337685](https://doi.org/10.1109/INFCOM.1994.337685)
18. Garcia-Luna-Aceves JJ, Madruga EL (1999) A multicast routing protocol for ad-hoc networks. In: Proceedings of the joint conference of the IEEE computer and communications societies (INFOCOM), vol 2, pp 784–792. doi:[10.1109/INFCOM.1999.751466](https://doi.org/10.1109/INFCOM.1999.751466)
19. Rubin I, Choi CW (1997) Impact of the location area structure on the performance of signaling channels in wireless cellular networks. IEEE Commun Mag 35(2):108–115. doi:[10.1109/35.565672](https://doi.org/10.1109/35.565672)
20. Zonoozi MM, Dassanayake P (1997) User mobility modeling and characterization of mobility patterns. IEEE J Selected Areas Commun 15(7):1239–1252. doi:[10.1109/49.622908](https://doi.org/10.1109/49.622908)
21. Johnson D, Maltz D (1996) Dynamic source routing in ad hoc wireless networks. In: Mobile computing. Kluwer Academic Publishers, pp 153–181
22. Broch J, Maltz D, Johnson D, Hu Y et al (1998) Multi-hop wireless ad hoc network routing protocols. In Proceedings of the ACM/IEEE international conference on mobile computing and networking (MOBICOM), pp 85–97
23. Chiang CC, Gerla M (1998) On-demand multicast in mobile wireless networks. In: Proceedings of the IEEE international conference on network protocols (ICNP), pp 262–270. doi:[10.1109/ICNP.1998.723747](https://doi.org/10.1109/ICNP.1998.723747)
24. Garcia-Luna-Aceves JJ, Spohn M (1999) Source-tree routing in wireless networks. In: Proceedings of the 7th international conference on network protocols (ICNP), p 273
25. Johansson P, Larsson T, Hedman N et al (1999) Routing protocols for mobile ad-hoc networks-a comparative performance analysis. In: Proceedings of the ACM/IEEE international conference on mobile computing and networking (MOBICOM), pp 195–206
26. Karlin S, Taylor HM (1975) A first course in stochastic processes, 2nd edn. Academic Press, pp 340–391 (Chapter 7)

27. Papoulis A (1991) Probability, random variables and stochastic processes, 3rd edn. McGraw-Hill
28. Rose C (1996) Minimizing the average cost of paging and registration: a timer-based method. *Wireless Netw* 2(2):109–116. doi:[10.1007/BF01225634](https://doi.org/10.1007/BF01225634)
29. Bai F, Sadagopan N, Helmy A (2003) IMPORTANT: A framework to systematically analyze the Impact of Mobility on Performance of Routing protocols for Adhoc Networks. In: INFOCOM 2003, twenty-second annual joint conference of the IEEE computer and communications, vol 2, pp 825–835. doi:[10.1109/INFCOM.2003.1208920](https://doi.org/10.1109/INFCOM.2003.1208920)
30. http://en.wikipedia.org/wiki/Manhattan_mobility_model. Accessed 24 Jan 2014
31. Akyildiz IF, Wang W (2002) A dynamic location management scheme for next-generation multitier PCS systems. *IEEE Trans Wireless Commun* 1(1):178–189. doi:[10.1109/7693.975456](https://doi.org/10.1109/7693.975456)
32. Bai F, Helmy A (2008) A survey of mobility models in wireless Ad hoc Networks. University of Southern California, U.S.A. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.107.7131> Accessed 26 Jan 2014
33. http://en.wikipedia.org/wiki/Renewal_theory. Accessed 24 Jan 2014
34. Li K (2011) Cost analysis and minimization of movement-based location management schemes in wireless communication networks: a renewal process approach. *Wireless Netw* 17 (4):1031–1053. doi:[10.1007/s11276-011-0332-0](https://doi.org/10.1007/s11276-011-0332-0)
35. <http://mathworld.wolfram.com/MarkovProcess.html>. Accessed 24 Jan 2014
36. http://en.wikipedia.org/wiki/Markov_process. Accessed 24 Jan 2014
37. Wang K, Huey J (1999) A cost effective distributed location management strategy for wireless networks. *Wireless Netw* 5(4):287–297. doi:[10.1023/A:1019163414359](https://doi.org/10.1023/A:1019163414359)
38. Akyildiz IF, Ho JSM (1995) A mobile user location update and paging mechanism under delay constraints. *ACM SIGCOMM Comput Commun Rev* 25(4):244–255. doi:[10.1145/217391.217455](https://doi.org/10.1145/217391.217455)
39. Casares-Giner V, Garcia-Escalante P (2004) On movement based mobility tracking strategy-a general framework. In: IEEE wireless communications and networking conference, WCNC, vol 4, pp 1957–1962. doi:[10.1109/WCNC.2004.1311385](https://doi.org/10.1109/WCNC.2004.1311385)
40. Chiang KH, Shenoy N (2004) A 2-D random-walk mobility model for location-management studies in wireless networks. *IEEE Trans Veh Technol* 53(2):413–424. doi:[10.1109/TVT.2004.823544](https://doi.org/10.1109/TVT.2004.823544)
41. Sheu TL, Wei WF (2010) A channel preemption model for vertical handoff in a WLAN-embedded cellular network. *Wireless Netw* 16(4):929–941. doi:[10.1007/s11276-009-0178-x](https://doi.org/10.1007/s11276-009-0178-x)
42. Baloch RA, Awan I, Min G (2010) A mathematical model for wireless channel allocation and handoff schemes. *Telecommun Syst* 45(4):275–287. doi:[10.1007/s11235-009-9267-5](https://doi.org/10.1007/s11235-009-9267-5)
43. Huang CJ, Lai WK, Wang CC et al (2001) A ratioed channel assignment scheme for initial and handoff calls in mobile cellular systems. *Comput Commun* 24(3):308–318
44. Trivedi KS, Dharmaraja S, Ma X (2002) Analytic modeling of handoffs in wireless cellular networks. *Inf Sci* 148(1):155–166
45. Seo JH, Han KJ (2004) Channel allocation scheme for handoff control in wireless Ad Hoc network with group mobility. *Telecommun Syst* 25(3–4):273–285. doi:[10.1023/B:TELS.0000014785.01627.22](https://doi.org/10.1023/B:TELS.0000014785.01627.22)
46. Sun C, Stevens-Navarro E, Shah-Mansouri V et al (2011) A constrained MDP-based vertical handoff decision algorithm for 4G heterogeneous wireless networks. *Wireless Netw* 17 (4):1063–1081. doi:[10.1007/s11276-011-0335-x](https://doi.org/10.1007/s11276-011-0335-x)
47. Chen Y, Chen H, Xie L et al (2011) A handoff decision algorithm in heterogeneous wireless networks with parallel transmission capability. In: IEEE vehicular technology conference (VTC Fall), pp 1–5. doi:[10.1109/VETECF.2011.6092894](https://doi.org/10.1109/VETECF.2011.6092894)
48. Zhang J, Chan HCB, Leung VCM (2006) Wlc14-6: a location-based vertical handoff decision algorithm for heterogeneous mobile networks. In: IEEE global telecommunications conference, GLOBECOM'06. IEEE, pp 1–5. doi:[10.1109/GLOCOM.2006.697](https://doi.org/10.1109/GLOCOM.2006.697)

49. Stevens-Navarro E, Wong VWS, Lin Y (2007) A vertical handoff decision algorithm for heterogeneous wireless networks. In: IEEE wireless communications and networking conference, WCNC 2007, pp 3199–3204. doi:[10.1109/WCNC.2007.590](https://doi.org/10.1109/WCNC.2007.590)
50. Song Q, Jamalipour A (2008) A quality of service negotiation-based vertical handoff decision scheme in heterogeneous wireless systems. *Eur J Oper Res* 191(3):1059–1074
51. Zhu L, Yu FR, Ning B (2010) An optimal handoff decision algorithm for communication-based train control (CBTC) systems. In: IEEE 72nd vehicular technology conference fall (VTC 2010-Fall), pp 1–5. doi:[10.1109/VETECF.2010.5594428](https://doi.org/10.1109/VETECF.2010.5594428)
52. Sundarapandian V (2009) Probability, statistics and queueing theory. PHI Learning Private Limited, New Delhi
53. Chung SP, Chien SC (2005) Analysis of multiservice cellular networks with asymmetrical traffic and handoff queue. *Comput Commun* 28(8):864–879
54. Louvros S, Pylarinos J, Kotsopoulos S (2007) Handoff multiple queue model in microcellular networks. *Comput Commun* 30(2):396–403
55. Sen SK, Bhattacharya A, Das SK (1999) A selective location update strategy for PCS users. *Wireless Netw* 5(5):313–326. doi:[10.1023/A:1019119632106](https://doi.org/10.1023/A:1019119632106)
56. Wu X, Mukherjee B, Bhargava B (2003) A low-cost, low-delay location update/paging scheme in hierarchical cellular networks. In: Proceedings of the 3rd ACM international workshop on data engineering for wireless and mobile access, pp 46–50. doi:[10.1145/940923.940932](https://doi.org/10.1145/940923.940932)
57. Munasinghe KS, Jamalipour A (2010) An analytical evaluation of mobility management in integrated WLAN-UMTS networks. *Comput Electr Eng* 36(4):735–751
58. Banerjee N, Wu W, Basu K et al (2004) Analysis of SIP-based mobility management in 4G wireless networks. *Comput Commun* 27(8):697–707
59. Banerjee N, Basu K, Das SK (2003) Hand-off delay analysis in SIP-based mobility management in wireless networks. In: Proceedings of IEEE parallel and distributed processing symposium, p 8. doi:[10.1109/IPDPS.2003.1213412](https://doi.org/10.1109/IPDPS.2003.1213412)
60. http://en.wikipedia.org/wiki/Multi-criteria_decision_analysis. Accessed 24 Jan 2014
61. Nasser N, Hasswa A, Hassanein H (2006) Handoffs in fourth generation heterogeneous networks. *IEEE Commun Mag* 44(10):96–103. doi:[10.1109/MCOM.2006.1710420](https://doi.org/10.1109/MCOM.2006.1710420)
62. Hasswa A, Nasser N, Hassanein H (2006) Tramcar: a context-aware cross-layer architecture for next generation heterogeneous wireless networks. In: IEEE international conference on communications, ICC'06, vol 1, pp 240–245. doi:[10.1109/ICC.2006.254734](https://doi.org/10.1109/ICC.2006.254734)
63. Hong KH, Lee SK, Kim LY et al (2009) Cost-based vertical handover decision algorithm for WWAN/WLAN integrated networks. *EURASIP J Wireless Commun Netw* 2009:15. doi:[10.1155/2009/372185](https://doi.org/10.1155/2009/372185)
64. http://en.wikipedia.org/wiki/Analytic_hierarchy_process. Accessed 24 Jan 2014
65. Saaty TL (1990) How to make a decision: the analytic hierarchy process. *Eur J Oper Res* 48(1):9–26
66. Song Q, Jamalipour A (2005) A network selection mechanism for next generation networks. In: IEEE international conference on communications, ICC 2005, vol 2, pp 1418–1422. doi:[10.1109/ICC.2005.1494578](https://doi.org/10.1109/ICC.2005.1494578)
67. Balasubramaniam S, Indulsk J (2003) Vertical handovers as adaption methods in pervasive systems. In: Icon 2003: 11th IEEE international conference on networks. The Institute of Electrical and Electronics Engineers, pp 705–710. doi:[10.1109/ICON.2003.1266273](https://doi.org/10.1109/ICON.2003.1266273)
68. <http://en.wikipedia.org/wiki/TOPSIS>. Accessed 24 Jan 2014
69. Kassar M, Kervella B, Pujolle G (2008) An overview of vertical handover decision strategies in heterogeneous wireless networks. *Comput Commun* 31(10):2607–2620
70. Zhang W (2004) Handover decision using fuzzy MADM in heterogeneous networks. In: Wireless communications and networking conference, WCNC 2004, vol 2, pp 653–658. doi:[10.1109/WCNC.2004.1311263](https://doi.org/10.1109/WCNC.2004.1311263)
71. http://en.wikipedia.org/wiki/Fuzzy_logic. Accessed 24 Jan 2014

72. Guo Q, Zhu J, Xu X (2005) An adaptive multi-criteria vertical handoff decision algorithm for radio heterogeneous network. In: IEEE international conference on communications, ICC 2005, vol 4, pp 2769–2773. doi:[10.1109/ICC.2005.1494852](https://doi.org/10.1109/ICC.2005.1494852)
73. Zekri M, Jouaber B, Zeghlache D (2010) Context aware vertical handover decision making in heterogeneous wireless networks. In: IEEE 35th conference on local computer networks (LCN), pp 764–768. doi:[10.1109/LCN.2010.5735809](https://doi.org/10.1109/LCN.2010.5735809)
74. Lee CH, Yu CJ (2004) An intelligent handoff algorithm for wireless communication systems using grey prediction and fuzzy decision system. In: IEEE international conference on networking, sensing and control, vol 1, pp 541–546. doi:[10.1109/ICNSC.2004.1297497](https://doi.org/10.1109/ICNSC.2004.1297497)
75. Edwards G, Kandel A, Sankar R (2000) Fuzzy handoff algorithms for wireless communication. *Fuzzy Sets Syst* 110(3):379–388
76. Ben-Mubarak MA, Ali BM, Noordin NK et al (2013) Fuzzy logic based self-adaptive handover algorithm for mobile WiMAX. *Wireless Pers Commun* 71(2):1421–1442
77. Honman B, Benjapolakul W (1998) A handover decision procedure for mobile telephone systems using fuzzy logic. In: The 1998 IEEE Asia-Pacific conference on circuits and systems, IEEE APCCAS 1998, pp 503–506. doi:[10.1109/APCCAS.1998.743835](https://doi.org/10.1109/APCCAS.1998.743835)
78. Homnan B, Benjapolakul W (2003) Adaptation of CDMA soft handoff thresholds using fuzzy inference system. *Wireless Pers Commun* 26(4):325–345. doi:[10.1023/A:1025684710889](https://doi.org/10.1023/A:1025684710889)
79. Homnan B, Benjapolakul W (2004) Application of fuzzy inference to CDMA soft handoff in mobile communication systems. *Fuzzy Sets Syst* 144(2):345–363
80. Zhu Y, Leung VCM (2006) A fuzzy distance-based location management scheme for PCS networks. In: IEEE 63rd vehicular technology conference, VTC 2006-Spring, vol 3, pp 1063–1067. doi:[10.1109/VETECS.2006.1682997](https://doi.org/10.1109/VETECS.2006.1682997)
81. Zhu Y, Yu L (2005) A dynamic forwarding pointer mobility management strategy with a fuzzy logic controller. *J Comput Res Dev* 42(12):2048–2055
82. Zhu Y, Pedryez W (2005) A fuzzy forwarding pointer location management strategy for personal communication networks. In: Proceedings of 2005 IEEE networking, sensing and control, pp 38–43. doi:[10.1109/ICNSC.2005.1461157](https://doi.org/10.1109/ICNSC.2005.1461157)
83. Taheri J, Zomaya AY, Iftikhar M (2011) Fuzzy online location management in mobile computing environments. *J Parallel Distrib Comput* 71(8):1142–1153

Part II
Classical Technologies and Applications

Chapter 5

Mobility Management at Link Layer

The typical wireless systems, including cellular mobile communication systems, Wireless Local Area Networks (WLANs) and satellite communication systems are briefly introduced first. Then, aspects of cellular systems including overview of network architecture, Mobility Management (MM) model and protocols, mobility management functions, and typical procedures are discussed. After that, security mechanisms of cellular systems from mobility point of view are summarized and compared. In the last section, interworking between WLAN and mobile communication is described.

5.1 Overview of Typical Wireless Systems

5.1.1 Cellular System

(1) History of Cellular System

The cellular system is the most widely deployed mobile communication system. Bell Laboratories in the USA first proposed the cellular mobile communication in the 1960s. The radio coverage area is divided into several cells. Each adjacent cell uses a different frequency to avoid cochannel interference, but the same frequency can be reused in non-adjacent cells. The Advanced Mobile Phone System (AMPS) was the first US cellular mobile communication system and became the fastest growing system since its inception in the 1980s. The AMPS is considered as one typical first-generation (1G) cellular system, which characterized by using analog voice technology and the frequency-division multiple access (FDMA) technology.

The first second-generation (2G) cellular system is deployed in the 1990s. The Global System for Mobile Communication (GSM) and code division multiple access (CDMA) are two typical 2G systems, both offering digital voice and short message services (SMSs), which adopt time division multiple access (TDMA) and

CDMA technologies at air interface, respectively. In the later stage of 2G (also known as 2.5G), data access service is offered by packet-switched domain (PS domain) at core network (CN), and with related function enhanced at radio network. Compared to the first generation, 2G systems have higher spectrum efficiency, enhanced roaming capabilities, improved security mechanisms, and begin to offer data access services.

For third-generation (3G) system, there are three different standards approved by ITU-T, i.e., wideband CDMA (WCDMA), CDMA2000, and time division-synchronous code division multiple access (TD-SCDMA). Compared to 2G systems, higher data rate and superior voice quality were achieved in 3G systems. The WCDMA and TD-SCDMA, both standardized by the 3rd Generation Partnership Project (3GPP), differ at air interface; however, their radio access networks (ANs) actually connect to same core network, i.e., a common core network consists of CS and PS domain. The first WCDMA system was deployed in 2001, and it has been used primarily in Europe, Japan, and China. In 2009, the first commercial TD-SCDMA network was deployed in China.

The Long-Term Evolution (LTE) and Long-Term Evolution-Advanced (LTE-A) standardized by 3GPP is one of the fourth-generation (4G) systems, and so does the Wireless MAN-Advanced developed (based on 802.16e) by IEEE. The LTE-A has a packet-switched core network-based IP technology, and the orthogonal frequency-division multiplexing (OFDM) and multiple-input multiple-output (MIMO) are two technologies used at radio interface to boost data rate. The LTE system provides almost seamless mobility for IP-based services between 2G/3G/LTE network and other radio access network such as WLAN, by utilizing network layer mobility management protocols developed by IETF with necessary optimization. The LTE networks (both TD LTE and LTE FDD, differ at radio layer) have been widely deployed across the world. The 4G offers us greater bandwidth, higher data rates, efficient spectrum use, and increased interoperability across the globe [1].

An overview of cellular systems is shown in Table 5.1.

(2) Basic Concepts of Cellular Systems

A. Network Architecture

From 1G to 4G systems, the names and functions of their network entities have been changed, accordingly. However, as shown in Fig. 5.1, the basic architecture, which consists of the mobile station (MS), the access network, and the core network, remains unchanged. The access network includes one or more wireless transceiver, also known as base station (BS), and a controller (Note: In LTE, related functions have been redistributed to eNodeB and core network entities.) who perform control function for those transceivers. The CN includes routing and switching equipment, a signaling control system (including call control and mobility management functions) and data repositories for subscription and charging. The interworking with other existing network is important from service provision point of view for each generation of cellular systems.

Table 5.1 An overview of cellular systems

	1G	2G/2.5G	3G	4G
First deployment year	1981	1991/1995	2001	2010
Main services	Analog voice	Digital voice, short message service; packet data services	Superior voice quality; broadband data;	Always-on data service, converged data and voice over IP
Typical standards	NMT, AMPS, TACS	GSM, CDMA; GPRS, EDGE	WCDMA, TD-SCDMA, CDMA2000	LTE-advanced, wireless MAN-advanced
Switching technologies	Circuit Switched	Circuit switched/packet switched	Circuit switched and packet Switched	Packet switched
Multiplexing	FDMA	TDMA, CDMA	CDMA	OFDMA
Duplex method	FDD	FDD	TDD FDD	FDD TDD
Data rates	–	9.6–14.4 kbit/s; 64–144 kbit/s	384 kbit/s–2 Mbit/s	100 Mbit/s–1 Gbit/s

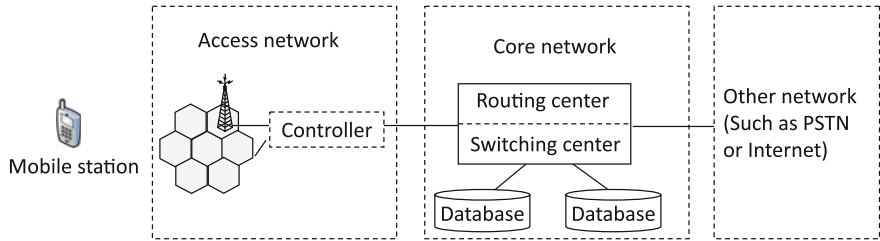


Fig. 5.1 Basic network structure for cellular systems

B. Radio Multiple Access technologies

Radio multiple access technology refers to techniques that enable multiple users to share the limited radio resources in an efficient and coordinated manner. As shown in Table 5.1, the FDMA and TDMA are used in 1G and 2G systems, respectively. The CDMA and OFDM are key technologies for 3G and 4G systems, respectively. The basic ideas of the above four technologies are briefed here, and readers interested in those technologies, please refer to [2].

- *FDMA* divides the available one frequency into several channels [2]. Each channel occupies a portion of the total available frequency band and is allocated to a single user when needed. Those channels are separated by guard bands (also a kind of channel with limited bandwidth, where no transmission allowed on such channels), thus radio interference is avoided.

- *TDMA* divides a radio frame into portions or time slots at the time axis direction, and each slot is assigned to a single user to transmit information [2]. The information from a single user always sits in the same time slot position of each frame.
- *CDMA* schemes based on spread spectrum technology offer a dynamic channel allocation radio multiple access schemes with no rigid channel allocation limitations for each individual user [2]. Using different uncorrelated codes, each user is able to send information on the same frequency spectrum without significant difficulty in detecting the desired signals at the receiver side. The CDMA is the basis for radio multiple access scheme of 3G radon interfaces.
- *Orthogonal Frequency-Division Multiplexing (OFDM)* is a frequency-division multiplexing scheme that uses a digital multi-carrier modulation method [1]. Large numbers of closely spaced orthogonal subcarriers are used to carry the information. The information is then divided into several parallel data streams or channels for each subcarrier.

C. Radio coverage technologies

In wireless mobile system before 1G (named as larger area system), each radio area covers a very large area, and this area is controlled by a base station using high transmitting power. Due to few radio channels in each area, the capacity of such system is limited, i.e., it only serves few users.

In order to serve more users, smaller coverage cells, frequency reuse, and automatically handover (i.e., no user intervention needed is detailed later this chapter) technologies have been introduced in cellular systems. The transmitting power of each cell is reduced greatly and it is well controlled to mitigate inference; thus, non-adjacent cells could use same frequency. In order to increase the system capacity further, one cell could be split into multiple smaller cells or sector cells (by using directional antenna). Theoretically, the cell is in hexagon shape, but the actual covered area is irregular.

The example radio coverage of large area system, cellular system, and cell splitting are shown in Fig. 5.2. The f letter followed by a numeric value denotes different frequency.

Cells can be categorized into femtocell, picocell, microcell, and macro-cell, according to their sizes [3]; their parameters and characteristics are shown in Table 5.2. In real network, different category cells are deployed to meet coverage and system capacity requirement, e.g., in dense populated area, more picocells could work at different frequencies, and macro-cell is used in rural area. Normally, radio coverage should be “seamless,” which means no “blind area” exists, i.e., radio signal is “good” enough everywhere within the coverage. Recently, lots of research [4, 5] focused on small cells, which is kind of cell usually covers a very small area to offload data traffic (some of those cells may only provide user plane to terminal).

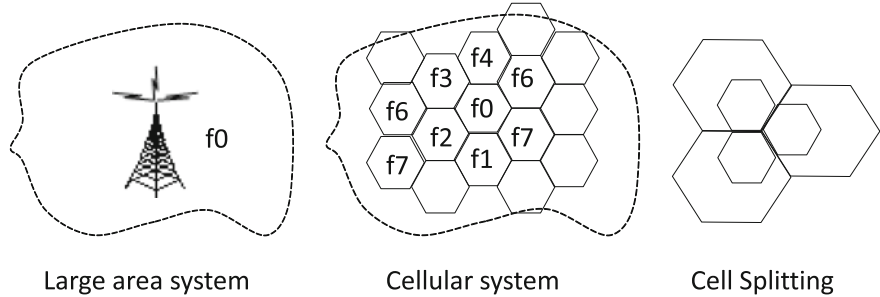


Fig. 5.2 The large area system, cellular system, and cell splitting from coverage aspects

Table 5.2 Category of cells

	Femtocell	Picocell	Microcell	Macro-cell
Coverage (in cell radius)	Up to 10 m or of the order of 10 m	Up to 500 m	500 m–2 km	1–20 km
RF power	Lowest	Low	Controlled by the base station	Usually high
Spatial/area spectral efficiency	High	High/medium	Medium	Low
Energy consumption	Lowest	Low	Medium	Highest
Target mobility speed	Low speed	Low and medium speed	Low, medium, and high speed	Low, medium, and high speed
Antenna aspects	Installed in door	Several radio antennae installed within building	Below average rooftop level	Taller than average roof-top level
Deployment scenarios	In home, hot spot,	Indoor coverage, e.g., small locale of airport, train stations, hot spots.	Urban areas	Urban and suburban areas
Special requirements for mobility management	Target: multiple cell types in coexistence scenarios Improved mobility performance, e.g., reduced handover failures, avoidance of ping-pongs, small cell discovery and e-ICIC			

5.1.2 WLAN

Within the WLAN, two or more devices communicate over radio media, usually at unlicensed spectrums, by wireless communication techniques (such as spread spectrum or OFDM). Most of WLANs deployed across the world are based on 802.11 series standards [6] proposed by the IEEE, such networks are referred as the “Wi-Fi.” The “Wi-Fi” and “WLAN” are interchangeable words in most cases.

The data access rate of Wi-Fi varies from several megabits/s to hundreds of megabits/s, while the latest standards (802.11ac) supports several thousands of megabits/s. In recent years, the study of IEEE is focusing on extending the range from hundreds of meters up to 1 km, and on how to utilize higher frequency band for WLAN, e.g., 5.8 and 60 GHz.

The Wi-Fi network normally consists of stations (STAs), access points (APs), and optionally access controller (AC), while the APs and ACs are connected via wired lines (i.e., a distribution system, DS). The infrastructure basic service set (IBSS) and extended service set (ESS) are two forms of Wi-Fi network infrastructure, which the former normally consist only one AP, while the latter are several BSSs connected via DS, which means multiple APs. The STAs may also work in “ad hoc” mode without any central point, which STAs communicate with each other directly, and this is also called independent basic service set (IBSS). The topologies of BSS, IBSS, and ESS are shown in Fig. 5.3, and the protocol stack of BSS is shown in Fig. 5.4. Lots of deployed Wi-Fi networks provide connection to the Internet.

When STA moves within one IBSS, there is no requirement defined for mobility. For the movement between APs which belong to the same ESS, the Inter-Access Point Protocol (i.e., IAPP, IEEE 802.11f) is used for accelerating the handover process, by link-layer mobility management method. For movement between APs which belong to different ESSs, the network layer mobility management method, e.g., Mobile IP, could be used. With the literature of Wi-Fi, the mobility between ESSs is named as “roaming.”

The speed of STA’s mobility is not considered when the first-generation Wi-Fi standards (i.e., IEEE 802.11a/b/g) were designed, since the target use case is fixed wireless indoor access, i.e., the STA is almost “fixed” at somewhere or move at

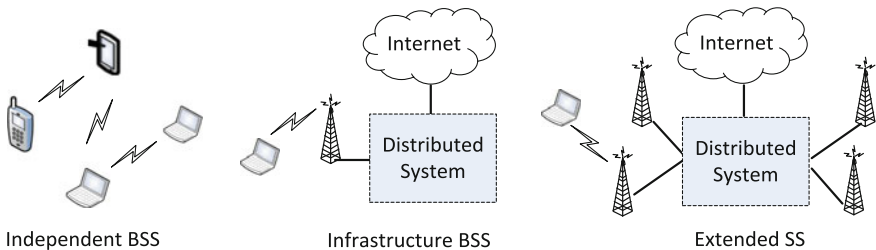


Fig. 5.3 Wi-Fi network, from topology perspective

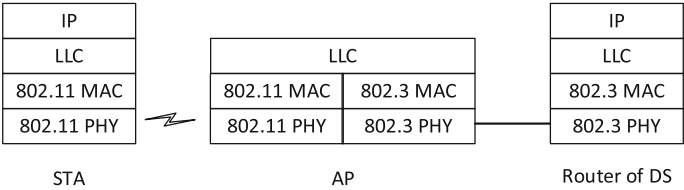


Fig. 5.4 Protocol stack of Wi-Fi, BSS

walking speed. Although the proposal targeting at speed around 60 km per hour has been brought up, it only receives little attention within the IEEE community.

One of the most recent research activities related to mobility management of Wi-Fi is how to enable autonomous global Wi-Fi roaming, i.e., when STA moves around, selection of Wi-Fi access network, authentication, and authorization is performed automatically without user involvement [7]. Another mobility management-related research activity is how to integrate WLAN into LTE network, which would be discussed in Sect. 5.7.

5.1.3 Satellite Communication Systems

Satellite communication systems are designed to provide communication services to users in under-populated regions, in emergency areas, on planes, trains, and ships, where terrestrial cellular systems and wireless systems are impossible to serve.

Those satellites working on different orbits around the Earth are categorized into geosynchronous Earth orbit (GEO) satellites and non-GEO satellites. As the name suggested, the GEO should maintain precise synchronization with the Earth’s rotation when it rotates. Those GEO satellites work at high frequencies and high altitudes (e.g., 35,800 km, such a significant distance that entails huge signal propagation delays and attenuation), so they are suitable for fixed satellite communication. Normally, those ground stations are equipped with large-size antennas, for high power transmit and facilitating receiving that requires high sensitivity [8]. In GEO system, no mobility happened from communication perspective, though those satellites rotate.

Non-GEO satellite system has two orbits: Low Earth orbit (LEO, which altitude is from 500 to 2000 km), and medium Earth orbit (MEO, which altitude is from 8000 to 12,000 km [8]). Compared with GEO system, the distance between the satellites and terminals or ground stations is shorter, thus required transmission power and propagation delays are reduced. Due to geosynchronous movement of the satellite is not maintained, the coverage of one non-GEO satellite moves continuously, even faster than that of a terminal at the Earth. For a given region to be served, once the satellite covers this region moves away, another satellite should move into provided radio coverage. This also requires the communication link

between one satellite and the terminal to be handovered from one satellite to another. Such kind of handover happens frequently and is also considered as link-layer mobility.

For GEO and non-GEO satellite communications systems, they have similar structures in general, i.e., consists of satellites, mobile terminals of various types, fixed ground stations (optionally), and gateway that connected to terrestrial networks, such as telephone network and the Internet. The major difference between those two systems is that normally GEO system has fewer satellites, typically three. However, the non-GEO system needs more satellites, to provide coverage continually for a give region regarding movement of satellites. Thus, most of those non-GEO systems need a constellation of satellites.

The Iridium system is the most famous LEO system of all those mobile satellite services (MSSs) today. It provides global radio coverage, including the poles, all the oceans, and the terrestrial surface, for voice and low bit-rate data transmissions service anywhere and anytime [9]. It has 66 active LEO satellites, connected with inter-satellite links (ISLs) among those satellites. Thus, a truly hubless airborne mesh network is constructed. An end-to-end path is established in the sky through those ISLs for voice calls and data communications, while connections via gateway (e.g., a ground station is needed) to other voice and data network for interworking are needed. From the movement perspective, the terminal is considered relative stationary compared with those orbiting satellites. If one satellite moves away, established radio link via this satellite, and related ISLs should be switched via suitable satellites, which is a kind of link-layer handover. Similar to cellular system, the mobile users home location, and visit location area are registered in the system, to support roaming among different countries. Details of the Iridium system and other MSS system could be found in [9].

Satellite communication systems still face lots of challenges, especially those MSS system. For example, how to provide seamless IP-based services by using network layer mobility method, e.g., Mobile IP or Proxy Mobile IP [8]. How can MSSs interworking with or integrated into a terrestrial broadband wireless system (e.g., LTE, Wi-Fi)? How to achieve seamless support of user mobility for satellite system? Some initial studies have already begun, but there remains lots of work to be done [8–10].

5.2 Overview of Network Architecture and Mobility Management in Cellular Systems

5.2.1 Network Architecture Evolution of Cellular Systems

Mobility Management technology is originated in cellular systems, together with radio transmission technologies; those two key technologies differentiate the cellular system with wired communication systems. One important principle of cellular

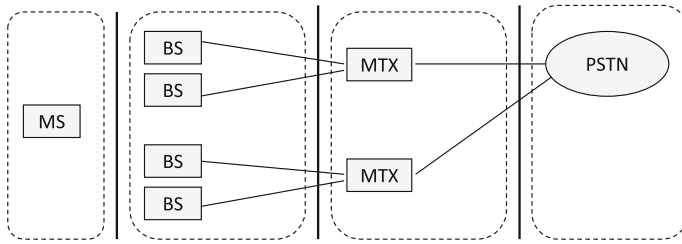


Fig. 5.5 Network architecture of NMT

system evolution is that the radio network and core network should evolve independently. The design work of 2G, 3G, and LTE systems comply with this principle. This principle gives more flexibility to the technical development and network deployment.

In this section, some typical systems from 1G to 4G are chosen for the introduction of network architecture and mobility management.

(1) Network Architecture and Mobility Management in NMT

The Nordic Mobile Telephone (NMT) system is a typical example of 1G, which consists of the mobile telephone exchange (MTX), the base station, and the MS (Refer to Fig. 5.5).

The MTX controls the traffic between the MSs in its operation area and the telephone network and also supervises the operation of its subordinated BS, via analog or digital signaling. The “roaming” is support, while it means if one MS has registered with its home MTX moves from a traffic area into a new traffic area, it automatically sets up an updating call to the visited MTX that in control of this new traffic area. Then, the visited MTX updates the MS’s home MTX with needed information, such as new “address.” There is no direct interface between two MTXs; thus, this updating is performed via the telephone network or other data network. The NMT also supports “switching call in progress” feature, where the MTX first orders surrounding BSs to perform signal strength measurements on the radio channel on which the MS is transmitting; then, it decides to which BS the call shall be transferred. The authentication procedure, which used to protect against illicit using of the subscriber identity number, takes place on all mobile originated calls; and optionally when roaming and on mobile terminating calls. The “scrambling” (security mechanism against eavesdropping on the radio link, which is optional) and short message service are also supported in NMT system. Details could be found in [11].

(2) Network Architecture and Mobility Management in GSM/GPRS System

The GSM is the most widely deployed 2G systems. Initially, the GSM system contains the MS, the BSS, and the NSS. The BSS provides and manages the transmission paths between the MS and the network. The base station controller (BSC) is responsible for radio resources management and handover control.

The NSS contains MSC/VLR, HLR/AuC, and GMSC. The MSC and VLR are always co-located in same physical entity, and so do the HLR and AuC. The GMSC connects GSM system to external telecommunication networks. At this stage, the GSM system is based on circuit switching technology, and circuit-switch-based services are only services provided.

In order to provide packet data service to user, GPRS technology has been developed. The PS domain, which contains SGSN and GGSN, is added at the network side. The Gb interface between SGSN and BSS is used to support packet data service. The GGSN works as a gateway, which connects GPRS network with external packet data network, such as the Internet.

Almost every entity of GSM/GPRS system contributes to the mobility management. The MS needs to be authenticated by core network, where the AuC, VLR, and SGSN are responsible for this procedure. For location management, the HLR, VLR, and SGSN store MSs' location, and VLR and SGSN are responsible for paging the MS, if there is a mobile-terminated communication for such MS. For voice and packet service in progress, the network entities from core network and BSS work in a coordinated way to handover MS (and its ongoing services) from source cell to target cell. The handover is "hard handover," i.e., the MS connects to either source cell or target cell during the handover procedure. The radio network resources for data plane at target cell should be established before issuing a "handover command" to the MS. The VLR and the SGSN also cooperate in some mobility management aspects. The "roaming" is supported in this way: where the MS is going to be served by a serving network (SN), and the serving network (VLR, SGSN) will request subscription data from user's home network (HLR) for performing user authentication and service authorization first. The serving network of a MS is the network it currently registered with, it is either one of those networks has roaming agreement with this user's network or users home network.

Figure 5.6 shows the GSM/GPRS network architecture, with entities and interfaces. Detailed information could be found on the 3GPP Web site.

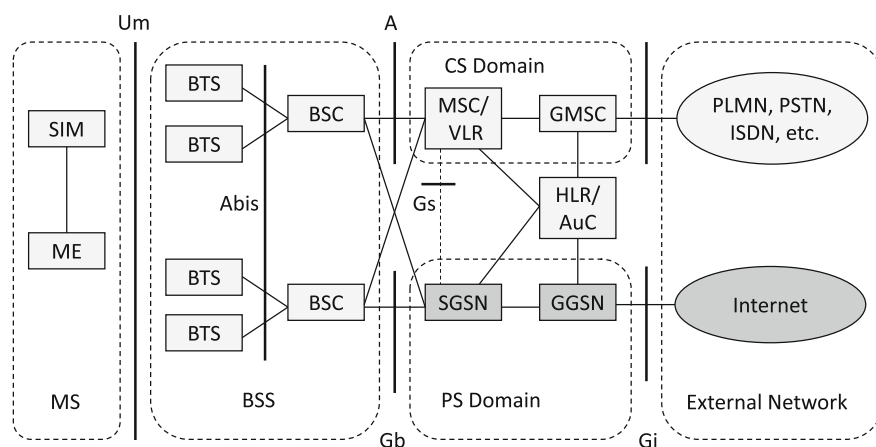


Fig. 5.6 GSM/GPRS network architecture

(3) Network Architecture and Mobility Management in UMTS System

The TD-SCDMA (Since Release 4) and WCDMA (since Release-99) are introduced in UMTS to improve spectrum efficiency, increase system capacity, and data access rate [12]. The network architecture of UMTS system is almost unchanged since its beginning. Within the UMTS radio network, it is the radio network controller (RNC) takes the full control of those NodeBs connected to it. The NodeB is the name for base station that operates TD-SCDMA or WCDMA air interface. From radio access technology’s (RAT’s) perspective, the “soft handover” and “baton handover” are supported for intra-WCDMA and intra-TD-SCDMA handovers, respectively.

However, the network evolvement is gradually and smoothly. The GSM/GPRS and UMTS network will coexist for a very long time for varies reasons, such as protecting investment, GSM/GPRS users unwilling to move to UMTS with some costs, and etc. The coexistence of varies radio access technology has brought up one important challenge for mobility management, that is, how to handle interradio access technology mobility? Corresponding methods were designed, both in the core network and radio access network. The fundamental of those methods lies in that the GSM/GPRS radio network are connected to the same UMTS core network, simultaneously as UMTS radio network does. Please refer to Fig. 5.7 for this architecture. The interface between RNCs (Iur) [13] and between RNC and BSC (Iur-g) [13] are also introduced for mobility management at the radio access network layer, i.e., between different radio networks.

During the evolvement of UMTS standards, some key features were introduced in each release [14]. For example, based on soft switch technology, the MSC has been divided to MSC Server and Media Gateway in Release-4. The IMS subsystem has been introduced to support voice over IP for UTRAN access in Release-5. In

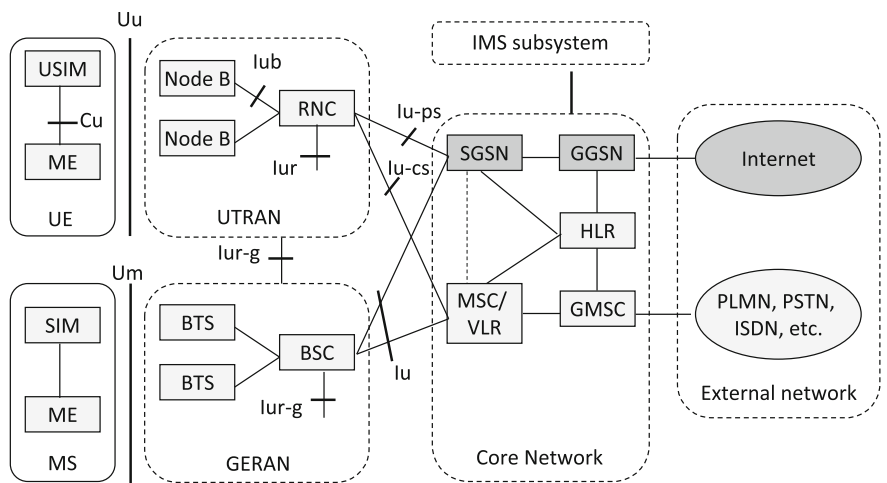


Fig. 5.7 UMTS network architecture

Release-4 and Release-5, underlying transport network begins to adopt ATM and IP. Above features have no impacts on mobility management.

However, the following features have impacts on mobility management. At the radio side, the radio access rate is increased by HSDPA and HSUPA, since Release-5 and Release-6, respectively. UMTS and WLAN interworking is standardized since Release-6. Since Release-8, “Home Node B” and new mobility management method are introduced. Mobility management aspects would be detailed later in this chapter.

(4) Network Architecture and Mobility Management in LTE system

Compared with GSM/GPRS and UMTS, the fundamental change of evolved packet system network (Release-8 is the first release [12, 14]) is that the EPS network is designed purely for packet data services. The EPS itself only has newly designed packet domain at core network side, and the air interface is also designed for IP packet only. The transport network of EPS is pure IP based, as well. From the point view of architecture, the EPS network has following characteristics: (1) a flat architecture which means fewer nodes was involved in user data traffic handling and (2) the control signaling and user data are handled separately by different entities. These allow the independent scaling, deployment, implementation, and optimization of control plane and user plane function.

Within the EPS core network, the Mobility Management Entity (MME) handles control signaling, while the Serving Gateway and PDN (Public Data Network) Gateway handle user plane data and bearer management-related signaling. The PDN-GW is a top-level anchor for achieving packet data service continuity; it is normally unchanged regardless user mobility, e.g., inter-RAT mobility or intra-RAT mobility. The Serving GW is a second-level anchor for user data tunnel and could be relocated due to mobility. The air interface of EPS network adopts OFDM and MIMO technologies [12], which boost the radio access rate as well as spectrum efficiency. It is named as LTE, and further divided into TD LTE and LTE FDD according to duplex method. The corresponding radio network is named as E-UTRAN, which only contains one element, the Evolved Node B (i.e., eNodeB). In order to interoperate with 2G/3G PS domain, the SGSN using S4 interface is the anchor for both signaling and user data to/from GERAN and UTRAN (if direct tunnel between UTRAN and SGW is not used). Another inter-operation option is to enhance MME and PDN-GW by adding current Gn and Gn/Gp interface functionalities, which would require no modification of deployed 2G/3G SGSN. Please refer to Fig. 5.8, for the basic EPS network architecture and one of its interoperation options [15, 16].

Although focused on providing “always-on” data connection for mobile Internet service, the EPS network still need to provide voice service for its registered terminals, such as to achieve voice service continuity for all mobility scenarios. There are two alternatives: circuit-switched fallback (i.e., CSFB) [17] and Single Radio Voice Call Continuity (i.e., SRVCC) [17]. The basic idea of CSFB is that when a voice call is either originated from or terminated for a UE when this UE is

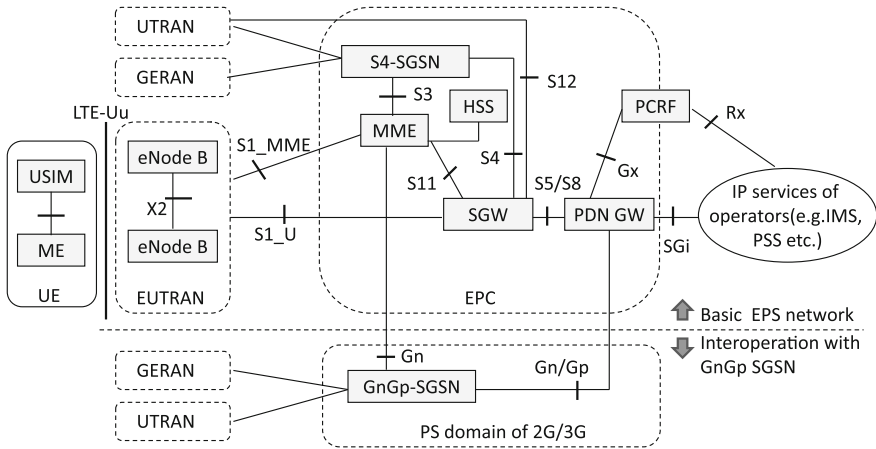


Fig. 5.8 EPS network architecture for 3GPP access and one of its interoperations

camping in LTE, the UE fallbacks to using existing CS domain to precede the voice call. The UE should reselect to LTE once the CS service is finished in CS domain. Within CSFB, the LTE, and 2G/3G radio coverage are overlapped. In order to coordinate mobility management functions between CS domain and EPS network, the SGs interface between MME and MSC Server is added to enable combined procedures for both EPS and CS service. And so does the S102 interface between the MME and the 1X CS interworking solution function (1X CS IWS) introduced for tunneling 1X CS service-related signaling to and from the UE camped in E-UTRAN and the 1X RTT MSC Server. Please refer to Fig. 5.9, for the network architecture for CSFB, where Media Gateway is not shown.

For SMS delivery, there are two options shown in the figure, SMS over SGs [17] and SMS in MME [17]. The short messages are piggybacked by signaling messages between UE and MME for both options. The MME chose one of the options depending on network configuration when UE registries to network. For the SMS over SGs option, short messages are transferred via SGs interface, for the SMS in MME option, the short messages are transferred via SGd interface, which is added mainly for in-bound roaming UEs whose home network is CDMA network [17].

The basic idea of SRVCC is that ongoing voice call is “handover” between IMS domain and CS domain when necessary, with only “single radio” interface activated on UE at a time to avoid terminal complexity brought by simultaneously operating two radio interfaces [18]. For voice call over IMS domain, all call-related signaling and data are transported via Packet domain to and from IMS domain. If the user moves to radio coverage with inefficient radio access data rate to support IMS voices, e.g., GERAN, then the ongoing voice call in IMS domain is “handover” to CS domain. Handling of the other leg of this voice call is not relevant to this “handover.” This direction of “handover” (i.e., from IMS to CS) is defined in

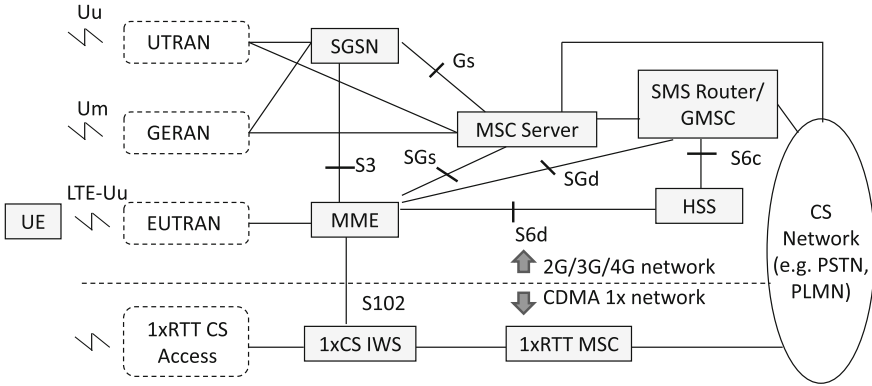


Fig. 5.9 Network architecture for CS fallback, and SMS in MME

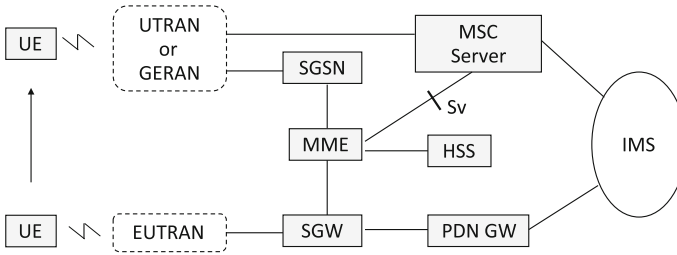


Fig. 5.10 Network architecture for SRVCC, from LTE to UTRAN/GERAN

Release-8. Please refer to Fig. 5.10. The SRVCC was enhanced in subsequent releases to support video services, and to reduce handover latency. Finally, in Release-11, the reverse SRVCC, i.e., the leg of ongoing voice call which is in CS domain being “handover” to IMS domain, is standardized.

From mobility management perspective, related network nodes are enhanced, and the Sv interface used for SRVCC signaling is introduced. Take the IMS to CS “handover” as an example; the SGSN and MME differentiate bearers used for voice from bearers used for other PS service during handover procedure. The MSC Server/MGW handles the bearer used for voice as a PS to CS handover procedure, and the SGSN will handle other PS bearers as a PS to PS handover procedure. For terminating voice call targeted to an idle-mode UE, the method of terminating access domain selection (T-ADS) is very critical to reduce call setup delay [19]. The goal of T-ADS is by selecting either IMS domain or CS domain, and selecting which type of radio network current terminal is possibility camped based information stored within network entities, to send paging message [18, 19].

The coexisting of multiple radio networks combined with requirement form service continuity has brought up lots of challenges to mobility management. The

mobility management solutions, methods used both at radio access network and core network become more and more complicated and will be discussed later.

The interworking between EPS and radio access technologies defined by other standard organizations, e.g., Wi-Fi, WiMax, and CDMA 1x eHRPD, has been studied since Release-8 [14, 20, 21]. The interworking architecture and mobility management aspects (some of them at network layer), are discussed together with WLAN-UMTS interworking in Sect. 5.7.

(5) Network Architecture and Mobility Management in CDMA2000

The first published standards of CDMA system is IS-95 series, which were defined by 3GPP2. The network architecture of IS-95 is similar to that of the GSM system. The CDMA2000 system is evolved based on IS-95, and the CDMA2000 1xEV-DO Release-0 is considered as one of the 3G systems. There are four major steps within the CDMA2000's core network architecture evolution: Phase 0, Phase 1, Phase 2, and Phase 3.

Phase 0 was the start point of the evolution toward an all-IP network. Compared with IS-95, the voice domain-related network entities and interfaces within this phase were almost unchanged for compatibility reasons, e.g., the transmission of A-interfaces is still based on circuit-switched technology. The network architecture and the mobility management mechanism of voice domain are also unchanged, i.e., still similar to those of 3GPP R99. The Wireless IP Network Architecture [22] from IETF is introduced to support wireless packet data networking. The following network entities are involved: the packet control function (PCF), the packet data serving node (PDSN), the home agent (HA), and AAA. In this phase, the air interface is also enhanced to support packet data services.

For packet data services, there are two kinds of services: the Simple IP service and the Mobile IP service. The Simple IP mode offers the basic packet data service, and its operation is similar to that of dial-up via modem. Firstly, the PPP session between the MS and the PDSN was established and the authentication succeeded; then, the PDSN (or cooperate with AAA, if needed) allocates an IP address to the MS. When the MS performs handover from one PCF to another PCF, the new PCF should first select the old PDSN, which currently serve the MS. If the old PDSN is selected, the RP session between the old PCF and the old PDSN was tore down and the new RP session between new PCF and the old PDSN is established to maintain the continuity of the establishes PPP session. If a new PDSN was selected and the MS still wants data services, a new PPP session would be established instead.

Based on the above link-layer mobility management used in the Simple IP services, the Mobile IP service is further enhanced to offers IP layer service continuity by using Mobile IP protocol and some extensions, e.g., the PDSN is enhanced with the foreign agent function; the MS should register with the home agent, etc. During the PCF to PCF handover, if the old PDSN could not be selected, then the new PCF would select a new PDSN. If new PDSN is selected, the PDSN to PDSN handover happens. If the Mobile IP registration were deemed required by the new PDSN, then the PPP session would be renegotiated. The change of foreign

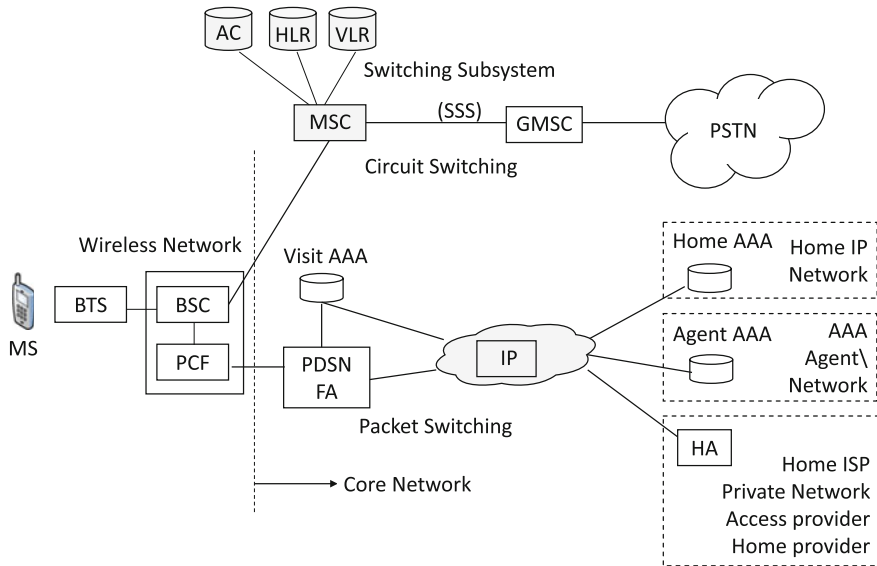


Fig. 5.11 CDMA2000 network reference model (with Mobile IP mode)

agent would be detected by the MS, and hence, the MS would reregister with its home agent. Details of Mobile IP Protocol and those extensions could be found on the IETF Web site.

The CDMA2000 network reference mode with Mobile IP mode is shown in Fig. 5.11. The corresponding protocol reference model is shown in Fig. 5.12.

Since Phase1, the radio network and core network begin to evolve independently. In CS domain, the MSC is divided into MSC emulation and the Media Gateway, and the ATM/IP is used for transmission. The above separation of signaling plane and media plane is similar to that of 3GPP R4. In the PS domain, the “always-on service” for MS’s PPP session is introduced, and the inter-PDSN handover latency is reduced by “fast handover” enhancement [22].

In Phase2, the “legacy mobile station domain (LMSD)” and the “IP Multimedia Domain (MMD)” are introduced. The LMSD can be roughly considered as the CS domain of the Phase2, with the difference that the LMSD using only IP-based transmission for radio and core network interfaces. The MMD offers both general packet data support capabilities and multimedia session capabilities, in which the latter is built on top of the former [23]. Roughly speaking, the MMD is the PS domain of the Phase2 network combined with the newly introduced IMS domain. The IMS domain is much similar to that of 3GPP R5.

Phase3 is the ultimate goal of the all-IP network. In this phase, the air interface would be IP-based, and the legacy MS will disappear. If supporting the legacy MSs became not needed, the LMSD would be not necessary and could be removed eventually.

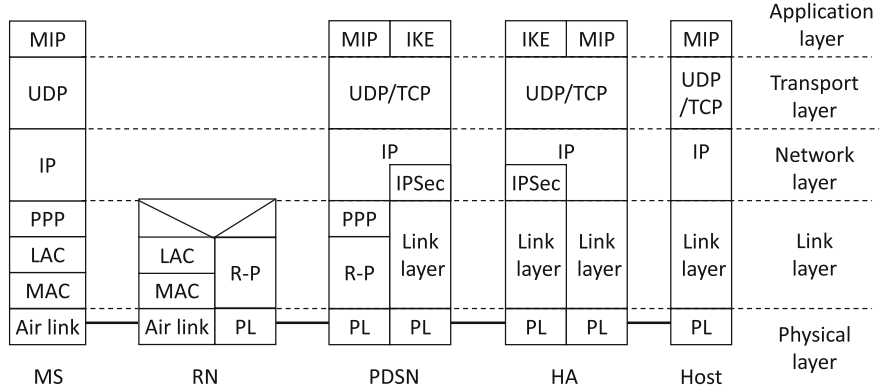


Fig. 5.12 Mobile IP-based CDMA2000 protocol reference model

5.2.2 Mobility Management Models of GSM/UMTS/LTE Systems

(1) Mobility Management Network Reference Model and Functions

As discussed in Sect. 2.2, the mobility management network reference model generalizes the mobility management functions, entities at an abstract level. Based on this reference model, mobility management entities and corresponding functions in GSM/UMTS/LTE are classified and summarized in Table 5.3. The security mechanism is intentionally not listed, since it has evolved dramatically from GSM to LTE. A brief introduction to such evolvments could be found in Sect. 5.6.

The mobility management functions of cellular system could be grouped roughly in four groups: radio access-related mobility management, location management, handover control, and security mechanism.

The radio access-related mobility management function includes those functions specific at the radio interface and within the radio network. Based on whether there is radio connection between MS and the radio access network or not, those functions could be further divided into mobility management in connected mode and in idle mode. For MS in connected mode, handover procedure related to radio layer (e.g., handover decision, preparation, and executions) and corresponding data forwarding. For MS entering the idle model, the cell selection or reselection should be performed by MS. The radio network also needs to scheduling and transmission of paging messages received from core network. The intra-RAT and inter-RAT scenarios are also considered by the above functions [24].

The location management refers to those functions mainly related to core network and non-radio access layer, for MSs have no connections with network, i.e., in idle mode. The registration, location-updating, and reachability management are three most important functions within the location management. Registration is the first step performed by the MS after it is powered on, during which the suitable

Table 5.3 Entities, functions in GSM/UMTS/LTE networks based on mobility management network reference model

Network reference model	GSM/GPRS	UMTS	LTE	Mobility management functions
Access points	BSS, BSC	Node B, RNC, home Node B	eNode B, home eNode B	Radio access-related mobility management
Domain mobile controller	MSC, SGSN and GGSN	MSC, SGSN and GGSN	MME, SGW and PGW	Handover control, including gateway and user plane anchor functions
Location server	VLR, and HLR	VLR and HLR	MME, HSS	Location management
Registration server	HLR, AuC	HLR, AuC	HSS, AuC	User subscription data management, location management

PLMN is selected, and the MS got authenticated and authorized by the network. The purpose of location updating is to keep the network updated where roughly the MS located and keep the MS registered as well. Where the MS is located is quit a large registered area contains many cells, where this area is denoted as location area, Routing Area and lots of tracking areas (TAs). The reachability management function is mainly used to find or “reach” the MS in idle mode, if there is a terminating call or data for it. The paging messages sent to each cells of MS’s registered area are used to notify the MS that there is terminating events for it. Such design of above three functions is a trade-off of costs between sending those paging messages and location-updating signaling, since per cell granularity location updating consumes too much network resources. Those location management functions used for each RAT, for CS domain, PS domain, and IMS domain are coordinated (e.g., idle-mode signaling reduction) to improve efficiency and performance.

The ***handover control*** refers to functions to maintain ongoing call or data session when user moves, within or between radio access technologies. Both the handover procedures happened at radio network and the handover control functions at core network side are integrated as a whole to ensure user’s voice and data session continuity when such mobility happens. Different types of handover according where it happens include CS handover, PS handover, which imply such handover happened within CS domain and PS domain. Handovers from PS to CS and from CS to PS procedures are specially used for SRVCC voice sessions. Different types of handover according to RAT involved including intra-RAT and inter-RAT handover, which imply whether such handover happens between the same RAT or not. Based on whether the domain mobile controller changes or not, handover type could be named as intra- and interhandover, such as intra-SGSN handover, inter-MME handover with or without Serving GW change, S1 handover, and X2 handover. From network-side perspective, the handover control function

prepares control plane and user plane resources at the target side before the actual handover happens at the radio network side and release resources at the source side after the actual handover finished. Data forwarding is used during handover procedure to achieve lossless handover, i.e., no user data loss during the handover, which is very important to data communications. The whole handover control functions are carefully designed to achieve the seamless handover, i.e., no service interruption is anticipated or aware by the user, which is very important for a voice call.

The ***security mechanism*** is the most important functions for any kind of mobile communication system. However, the radio interface and user mobility have brought additional security challenges to the mobile system. Basically, user authentication and service authorization are performed for every user and its MS based pre-shared security credentials between the SIM/USIM card and the AuC, except for emergency service in which such procedure could be ignored according to rules. From the user perspective, protection for user identity, data and location privacy is another important goal. From the network operator side, radio and core networks also need to be protected against eavesdrop, resource abuse, malicious attack, and etc. Solutions such as ciphering and integrity protection functions were introduced in GSM/GPRS system firstly and further enhanced in UMTS and LTE system, as well as other security methods.

(2) Protocols Associated with Mobility Management

As discussed in Sect. 2.1, mobility management protocol reference model generalize the mobility management-associated protocols, for data plane, control plane, and management plane, respectively. Based on this protocol reference model, several protocols that have mobility management-associated functions are selected from GSM/UMTS/LTE and are summarized in Tables 5.4 and 5.5, for data plane and control plane [25–27], respectively.

At air interfaces, protocols are divided into access stratum (AS) and non-access stratum (NAS). Within AS, such protocols terminate between the MS and the radio network node. The RR protocol and RRC protocol perform most critical control functions for radio interfaces. For example, the RRC in LTE has following functions [24]: (1) broadcast of system information; (2) paging; (3) establishment, maintenance, and release RRC links and radio bearers between MS and eNodeB; (4) security-control-related functions; (5) radio link control (RLC); and (6) mobility management functions, such as MS measurement reporting and control, handover control, and MS cell selection and reselections. In LTE and UMTS, the packet data convergence protocol (PDCP), radio link control, and medium access control (MAC) are considered as data plane protocols or sublayers of AS, with little difference in each system. The PDCP has transmission functionality that maps characteristics of higher-level protocols onto the characteristics of the underlying radio interface protocols. For upper layer data, the PDCP provides header compression, ciphering, in-sequence delivery of upper layer data. The RLC provides a radio-solution-dependent reliable link, e.g., concatenation, segmentation, and

Table 5.4 Protocols in GSM/UMTS/LTE networks for data plane

	GSM/GPRS	UMTS	LTE
Protocol used at air interface (access stratum)	LAPDm in early GSM release, evolved to use PDCP, RLC, MAC since Rel-4	PDCP, RLC, MAC	PDCP, RLC, MAC
Protocol used at air interface (non-access stratum)	RLP(for CS domain, between MS and MSC); SMDCP, LLC (For PS domain, between MS and SGSN, for A/Gb mode)	–	–
Protocol used at RAN and CN border	Framing protocols, such as ATM, Frame Relay; evolved to use IP in later release; BSSGP for A/Gb mode; GTP-U for Iu mode	GTP-U (For PS domain); framing protocols or IP in later release to transport voice packets	GTP-U
Protocol used between CN nodes	Framing protocols; evolved to use IP in later release when soft switch introduced for CS domain; GTP-U for PS domain	GTP-U	GTP-U

Table 5.5 Protocols in GSM/UMTS/LTE networks for control plane

	GSM/GPRS	UMTS	LTE
Protocol used at air interface (access stratum)	RR in early GSM releases, evolved to use RRC since Rel-4;	RRC	RRC
Protocols used at air interface (non-access stratum)	CM, MM (between MS and MSC); GMM, SM (between MS and SGSN)	CM, MM (between MS and MSC); GMM, SM (between MS and SGSN)	EMM, ESM (between UE and MME)
Protocol used at RAN and CN border	BSSAP for A/Gb mode; RANS-AP for Iu mode	RANAP	S1-AP
Protocol used between CN nodes	MAP; GTP-C	MAP, diameter; GTP-C	Diameter; GTP-C

reassembly RLC SDUs. The MAC controls the access signaling procedures for the radio channel, e.g., logical channels multiplexing, mapping between logical channels and transport channels, and physical channel mapping. However, they also provide underlying transport services for RRC signaling messages and associated functions.

Within NAS, protocols terminate between the MS and the core network node. There are several control plane function entities or sublayers of NAS [28, 29], such

as connection management (CM), mobility management for CS domain, session management (SM), GPRS mobility management (GMM) for PS domain, and EPS mobility management (EMM), and EPS session management for EPS network. For CS domain, the MM performs the mobility management functionality for non-GPRS service, such as IMSI attach, detach, and security-related function such as user identity confidential protection and location area update. The CM performs functions related to CS domain voice call control, supplementary service management and the short message service. It manages the establishment and release of voice call connections between mobile users. For PS domain, the GMM performs the mobility management functionality, such as GPRS attach, detach, security, and routing area update. The SM performs user plane-related function, e.g., PDP context activation, and PDP context deactivation. The EMM and ESM are EPS system counter parts of GMM and SM. But with enhancement such as tracking area update could be triggered based on user history mobility patterns. The ESM, EMM, and relation functions work together to support the “always-on” IP connection, which is a distinguished LTE feature for mobile Internet applications. Only in the A/Gb mode, that has two user plane protocols between MS and SGSN: (1) Subnetwork Dependent Convergence Protocol (SNDCP), and (2) logical link control (LLC), which provide data plane functions much similar to that provided by PDCP and RLC in LTE and UMTS.

At the RAN and CN border, the BSSAP, Radio Access Network Application Protocol (RANAP), and S1-AP are used as control plane protocol for A/Gb, Iu, and S1 interfaces. The major function of those control plane protocol are encapsulates and carries higher-layer signaling, handles signaling between the RAN node and CN node, such as between 3G-SGSN and RNC for RANAP, and manages the user plane tunnels such as GTP-U at the border. For user plane with CS domain, i.e., A/Iu-CS interface, it initially uses framing protocol to transport voice frame and later turned to use IP. For PS domain, the Gb interface uses the BSSGP protocol. The Iu-PS and S1 interface both use GTP-U version 0 [30]. Those protocols are used to tunnel user packet data between RAN and CN node, as well as between CN nodes (e.g., between SGW and PGW).

Within the core network, the CS domain initially designed based on SS7 to support voice service, e.g., via the MAP protocol, to support mobility management functions, such as user registration, MS location reporting signaling between MSC/VLR and HLR, handover control signaling between different MSC/VLRs, and voice call-related signaling. The MAP is also used within PS domain of UMTS (e.g., between SGSN and HLR), for location management and security management purposes. While within EPS core network, mobility management-related protocol procedures are implemented by enhancing the diameter protocol, which is an IP-based protocol. For the PS domain and EPS, the GTP-C [31] is used to activate/deactivate/modify PDP context in UMTS and create/delete/modify EPS bearers, respectively. The GTP-C also performs handover control-related functions for PDP context and EPS bearers, such as creating new bearers at target side and

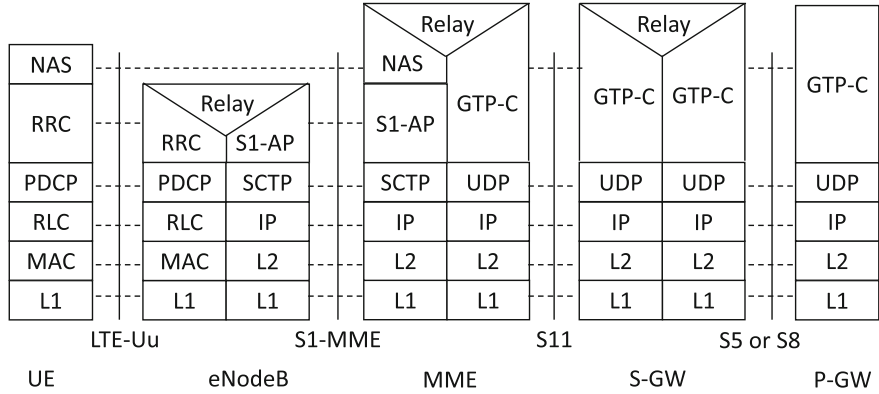


Fig. 5.13 Control plane protocol stacks between UE and P-GW (with E-UTRAN)

removing original bearers at source side, user mobility, and PDP/bearer context transfer between target and source side node. For EPS core network, the PMIP protocol is an alternative for GTP used only between SGW and PGW (i.e., PMIP variant S5/S8 interface).

For better understanding above protocols and their associated functional entities, control plane and user plane protocol stacks between UE and P-GW of LTE system [15] are taken as examples in Figs. 5.13 and 5.14.

The distributions of mobility management-related functions among network entities and related protocols are shown in Fig. 5.15.

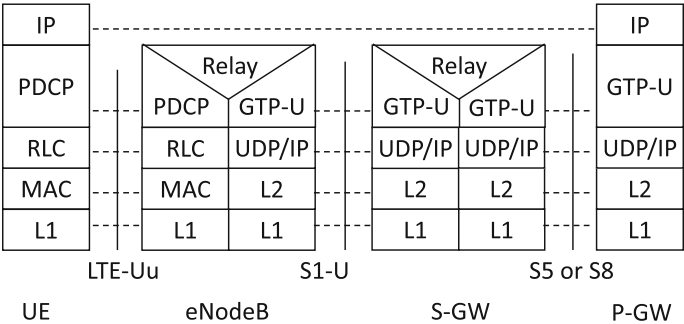


Fig. 5.14 User plane protocol stacks between UE and P-GW (with E-UTRAN). Reprinted from 3GPP TS23.401 [15]

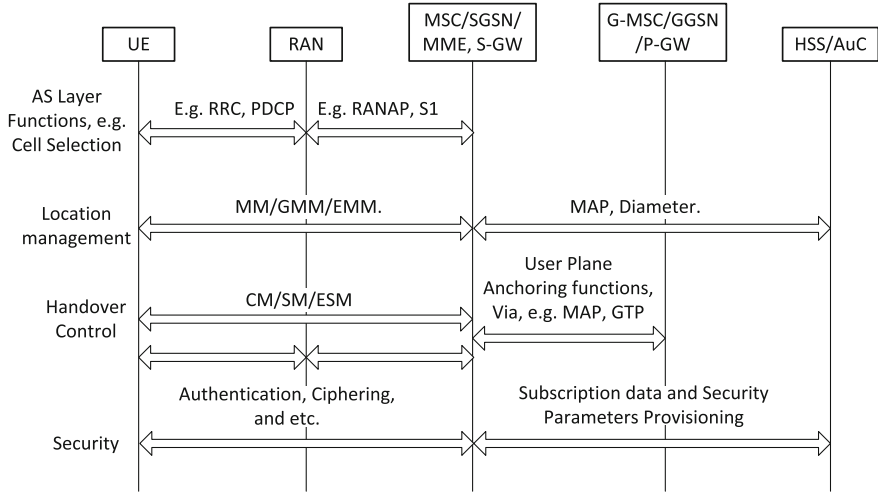


Fig. 5.15 Distributions of mobility management functions

5.3 Mobility Management Within Radio Access Networks

The separation of RAN and CN allows both parts could evolve independently, that is, matured wireless technologies could be adopted as soon as possible on RAN side without impacting on core network. The mobility management aspects more specific to radio access network is introduced in this section; then, system discussions for location management and handover control are also discussed in the next two sections.

5.3.1 Mobility Management in the Idle Mode

If there are no data to be sent or received, the MS would leave connected mode (also referred as active mode), and goes into the idle mode, i.e., there no connections between MS and the network since such connections have been released. When a MS powers on, it also enters into the idle mode firstly. Within the idle mode, the PLMN selection, cell selection, and reselection, location registration, and the supports for manual CSG cell selection are four major tasks from UE point of view [31]. Please refer to Fig. 5.16.

Those four tasks are all coordinate processes between NAS and AS layer. For *PLMN selection*, the NAS selects a PLMN either automatically by utilizing information stored on the SIM card related to PLMN selection, e.g., a list of priority ordered entries, each contains a PLMN and one or more radio access technologies [32], or based on user's manual input (normally, AS layer scans and reports available PLMNs to NAS, autonomously or upon requested by NAS; then, either

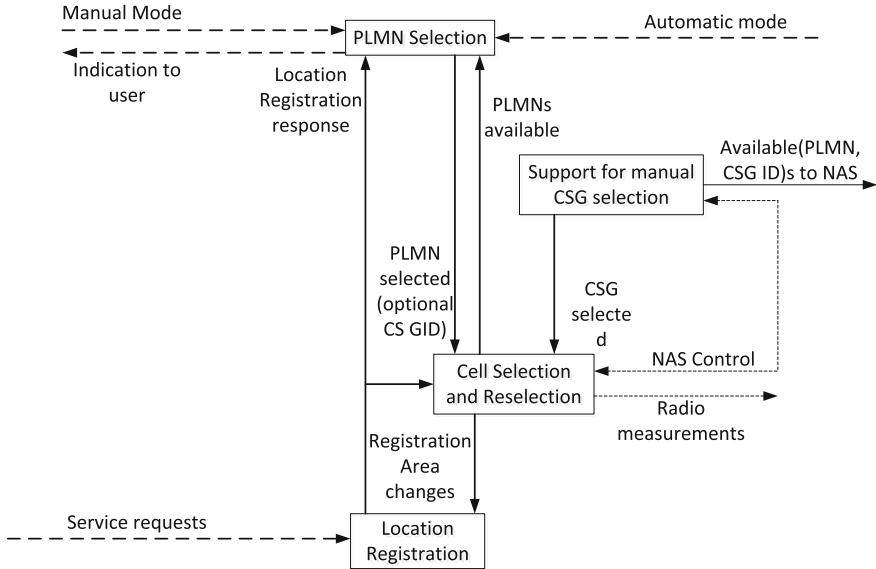


Fig. 5.16 Overall idle-mode process. Reprinted from 3GPP TS36.300 [24]

those PLMNs are displayed on screen for user to choose or NAS select one on behalf of user). Then, NAS requests the AS to select a cell (i.e., *cell selection*) that belongs to this selected PLMN. The NAS can also control in which RAT(s) the cell selection should be performed, by indicating those RAT(s) associated with the selected PLMN to the AS layer. If such information is not provided, the cell selection is performed within the RAT(s) that the UE supported.

By detecting and synchronizing to a broadcast channel, the UE finds one or more “PLMN identity” from the system information. Then, AS layer searches and selects a suitable cell if the idle-mode measurements result of a cell meets the cell selection criteria, e.g., for LTE, the measured cell receive level value and quality value meet that they are both greater than zero [33]. Once a suitable cell is found, the UE chooses this cell to provide available services, and further, the UE tunes to this cell’s control channel. This choosing is known as “*camping on the cell*” [33]. When camped on a cell, the UE is able to receive system information from the PLMN, e.g., registration area information such as tracking area identity of this cell and transfers to connected mode (i.e., establish radio connections) if it has registered with the PLMN, receiving “paging” messages sent on control channel, receiving ETWS and CMAS notifications, and receiving MBMS service.

The *location registration* is the next procedure if UE camped on a suitable cell after it powered on. The UE registers its presence, by means of a NAS registration procedure (e.g., attach procedure) to the selected PLMN. After a successful registration, the selected PLMN becomes the registered PLMN, and the registration area where this procedure happened became the registered registration area of this UE.

When camped on a cell, the UE shall regularly search for a better cell according to the cell reselection criteria. Priorities between different radio frequencies or RATs may be provided to the UE by system information or dedicated signaling, and priorities-related criteria (e.g., E-UTRAN inter-frequency and inter-RAT cell reselection criteria) may be considered additionally in cell reselection process. By setting suitable priorities, load balancing among different frequencies/RATs could be achieved at the radio layer. Thus, the reselected better cell may belong to another frequency/RAT which is different from that of current camped cell [33]. Depending on UE's mobility state, e.g., normal-mobility, high-mobility, and medium-mobility states, some parameters are scaled to make the reselection more reasonable. For example, either in high-mobility or medium-mobility state, the hysteresis value is adjusted accordingly for avoiding "ping-pong" reselections as much as possible. The mobility state is decided on how many times the reselection that UE performs during a fixed period, rather than its actual mobility speed.

When UE is leaving connected mode, the Redirected Carrier Frequency maybe provides by network to UE, and it is used to redirect the UE to an E-UTRA; or an inter-RAT carrier frequency is to be used by cell selection procedure upon leaving the RRC_CONNECTED [33].

The NAS is informed if the cell (re)selection results in changes in the received system information relevant for NAS compared to previously system information which is stored locally, e.g., if the registration area is changed, which may trigger NAS to initiate the Location Registration procedures. Since having been registered, the UE initiates location update or routing area update, or tracking area update procedure accordingly, which depends on the RAT of the reselected cell and the service (e.g., CS service or PS service or both), to update its presence to its registered PLMN.

The closed subscriber group (CSG) cell is a cell that broadcasting a specific CSG identity and only provide access to users belongs to the CSG identified by the same identity if such CSG cells operate in close mode. Normally, a cell operated by home eNodeB or home node (i.e., H(e)NB) acts as a CSG cell, for a specific group of user, e.g., the owner of H(e)NB. CSG cells are candidate cells as well as other non-CSG cells for normal cell selection and reselection process, if the UE has the permission to access such CSG cells. The manual CSG cell selection allows the end user to select a CSG and an associated PLMN. The AS first scans and provides available CSG cells' CSG identities and associated PLMNs to NAS, then such identities would be displayed to user. Based on user's selection of a CSG, NAS requests AS to select a cell belonging to this CSG.

5.3.2 Mobility Management in the Connected Mode

The UE has registered successfully with the network enters into the connected mode, if there is data/signaling to be sent or it needs to answer received paging messages in the idle mode, e.g., by initiating service request procedure via a current

camped LTE cell. During this service request procedure, both the radio links between UE and base stations and those under layering tunnels (e.g., GTP-U tunnels) within wired core networks used to transport user connection (e.g., a voice call connection or an IP session) should be maintained until user connection is terminated. It is radio network's responsibility to supervise whether there is data to be transferred between the terminal and the network. During a long period, if no data transferring has been detected for a terminal which is in connected mode, the radio network requests the core network to release control and user plane connections for this terminal via, e.g., sending Iu-Release or S1-AP UE Context Release Request to network. The core network may also initiate such release procedure due to some causes, such as authentication failure, and detach, to push terminal enter to idle mode. Radio links are released after radio resource control release procedure is completed. After that, the terminal enters the idle mode.

Due to user mobility, or variance of radio propagation or network load balancing reasons, current radio links may become not able to serve this UE anymore. In such a case, user connections should be transferred to another radio link. The transfer of a user's connection from one radio link to another is named as Handover [34]. During the handover process, both the ongoing service and its associated quality of service should be maintained, i.e., without any perceptible service interruption and degradation. To achieve higher handover successful rate and lower handover delay are two major goals of mobility management in connected mode from whole system point of view. At the same time, the signaling cost incurred by handover should be minimized as well.

However, handover may not always possible in the case that target system (e.g., 2G GPRS) does not support inter-RAT handover from a different type of source system (e.g., LTE) due to higher cost and huge impacts incurred by upgrading such target system to support inter-RAT handover. In such a case, cell change order (CCO) or redirection is used to move the terminal to target radio access, with relative lower cost but hundreds of milliseconds up to one- or two-second delay.

In general, cellular systems with smaller cell sizes require faster and more reliable link quality evaluation and handover algorithms. Labedz [35] has shown that the number of cell boundary crossing is inversely proportional to the cell size. Furthermore, Nanda [36] has shown that the handover rate increases with only the square root of the call density in macro-cells, but linearly with the call density in microcells. Since the mobile terminal has a certain probability of handover failure each time a handover is attempted, it is clear that handover algorithms must become more robust and reliable as the cell sizes decrease [37].

In the rest of this section, different categories of handover, an example handover procedure without core network involved, methods used to move connect mode terminal from one radio access to another, i.e., CCO and redirection are introduced.

(1) Handover Categories Based on Which Entity Decides to Handover

The handover measurement, decision, and execution are three steps of a handover procedure. Based on which node makes the final handover decision,

handover could be categorized as mobile-controlled handover (MCHO) and network-controlled handover (NCHO). In MCHO, it is MS's responsibility that continuously measuring radio link quality, and if handover criteria is met, the MS chooses the best radio channel then imitates a handover request to network. In NCHO, the base station measures qualities of radio channels that MSs are using continuously, and reports results to the mobile switch center. The MS center chooses new radio channel which is good enough for the MS, and controls the handover process, i.e., command MS to switch to using new radio channel.

The Mobile-assisted handover (MAHO) is considered as a kind of evolvement of NCHO, where instead of the base station, it is a MS that measures radio channel and reports results to the network. However, how to do the measurement is under networks controls. It is also called mobile-assisted network-controlled handover, and it is the principle that followed by handover design of GSM, CDMA, WCDMA, TD-SCDMA, and LTE system. We take intra-LTE system handover as an example, to illustrate the radio network aspects of such mobile-assisted network-controlled handover procedure.

Figure 5.17 depicts one handover procedure during which MME and Serving GW are not changed, i.e., intra-MME/S-GW handover, also known as X2 handover without S-GW relocation [24, 38].

The measurement object, report configuration, measurement gap, measurement value configurations, threshold to start the measurement, and speed-related parameters are included in measurement configuration information, which is provided by eNodeB to UE via RRC message. Frequencies of different RAT to be measured are included in measurement object. The periodical report and events trigger reporting configuration are included in the report configuration.

Once the quality of the serving cell signal decreased and reached the threshold, UE would start measurement according to the received configuration. If the threshold is not configured, UE performs mobility measurement continuously. For each cell that belongs to GSM, UTRAN, and LTE, different parameters are measured and used to evaluate the quality of measured cell. Readers interested in such physical layer procedures are encouraged to refer to other books for details.

The measurement gap is used for UE who is not capable of measuring and transferring data at the serving cell simultaneously. The UE measures neighbor cells at the time period (normally 6–8 ms) calculated according to this Gap, while during this time serving cell would not schedule any transfer for this UE.

The measurement reporting is triggered if reporting condition is met, e.g., configured event has been detected, or reporting comply with the periodical reporting configuration. For example, once event A4 (Neighbor becomes better than threshold) is detected, UE should report measurement results to the network. Events A1–A7 are defined for intra-LTE and Events B1–B2 are defined for inter-RAT case in [38] for LTE.

In step 4, depending on measurement report received from the UE and local information, the source eNodeB selects a target eNodeB and decides an X2 Handover would be performed (following requirements should be met: no MME change will happen, X2 interface is present between source and target eNodeB, and

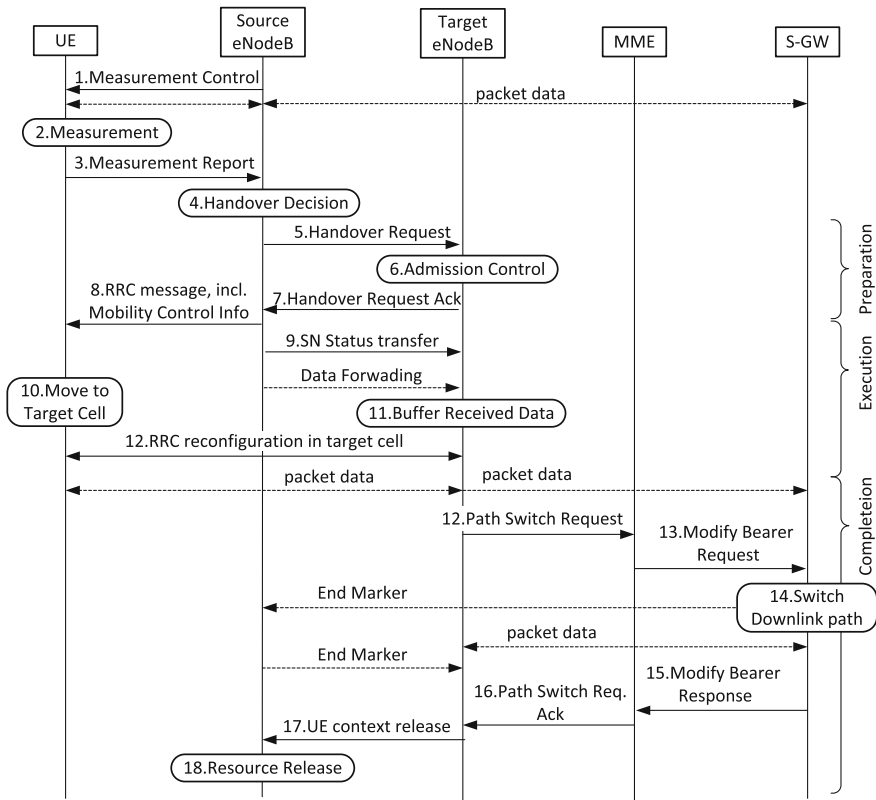


Fig. 5.17 Intra-MME/S-GW handover procedure. Reprinted from 3GPP TS36.300 [24]

source and target eNodeB both have IP connections with Serving GWs; otherwise, the eNodeB would initiate an S1 handover, and please refer to Sect. 5.5.2), and then it sends handover request which including necessary information such as *UE X2 and S1 signaling references* and evolved radio access bearer (E-RAB) context, to the target eNodeB. The *signaling references* enable the target eNodeB to address the source eNodeB and the EPC node. The *E-RAB context* includes necessary radio network layer and transport network layers addressing information, and quality of service (QoS) profiles of the E-RABs.

Depending on E-RAB context information, the target eNodeB perform admission control (step 6) sees whether required resources can be granted (this would increase the likelihood of a successful handover), and configures E-RAB-related resources at radio access stratum, and prepares other necessary access stratum sources. Within the handover request acknowledge message sent to source eNodeB, the target eNodeB includes a transparent container, an RRC message will be sent to the UE by source eNodeB as to perform the handover. The container includes a new C-RNTI, target eNodeB security algorithm identifiers for the selected security

algorithms, may include a dedicated RACH preamble, and possibly some other parameters, i.e., access parameters and system information blocks (SIBs), which would help terminal access at the target cell faster [24].

For data plane handling, the source eNodeB first decides which E-RABs are subject to “lossless handover,” i.e., no packet loss and in-sequence delivery are both guaranteed for those bearers during handover, and which E-RABs are not. During preparation phase, direct forwarding tunnels (one for uplink, another one for downlink) between source and target eNodeB were established. As soon as the source eNodeB receives handover request acknowledgment or transmitted handover command to terminal, source eNodeB starts data forwarding for those E-RABs. The source eNodeB sends the SN status transfer message to the target eNodeB to convey following information for those bearers that need PDCP status preservation (i.e., lossless and ordered PDCP service data delivery, please refer to [24] for details of PDCP status preservation), such as the PDCP SN of the first missing Uplink SDU, out of sequence Uplink SDUs that the terminal needs to retransmit in the target cell, the next PDCP SN that the target eNodeB shall assign to new downlink SDUs, and etc. The target eNodeB buffers received forward data.

In steps 10–11, the terminal moves to target cell and completes the radio connection configuration in this cell for all its bearers. After that, target eNodeB sends path switch request to request switch downlink data path to target eNodeB and starts forward buffered data to terminal. At the same time, the terminal is allowed to send uplink data to Serving GW. Upon the receipt of Modify Bearer Request, Serving GW sends one or more end marker packets on old path after it switches downlink path to target eNodeB. The end mark packets are used in PDCP status preservation function, by target eNodeB.

In steps 15, 16 and 17, source eNodeB is notified that downlink path switching is succeeded, and then resources related to this terminal are released.

(2) Handover Categories Based on Other Criteria

A. Hard Versus Soft Versus Baton Handover

Handovers may be categorized into hard handover [39], soft handover and baton handover, based on how radio links are handled during the handover procedures. In hard handover, all the old radio links in the mobile device are abandoned before the new radio links are established. The soft handover is quite different, where the radio links are added and abandoned in such a manner that the mobile device always keeps at least one radio link to the UTRAN, that is, the mobile device could have both radio links via old cell and via new cell [40]. The baton handover, as the name suggest, the uplink radio connections are firstly handed over from the old cell to the new cell before handing over downlink radio link, and the mobile device can send uplink data via new cell while receive downlink data from downlink radio connection connected via old cell during the handover procedure. The hard handover is widely used within GSM and LTE radio network, the soft handover is used within WCDMA radio network, and the baton handover [41] is used within TD-SCDMA radio network.

B. Intra-RAT Versus Inter-RAT Handover

Handovers may be categorized into intra-RAT handover, inter-RAT handover, based on whether the radio access technology used before and after the handover is different or not. If same, then this handover is intra-RAT handover; otherwise, it is inter-RAT handover. The inter-RAT handover is also regarded as inter-system handover, which means that the handover is between networks using different radio systems, e.g., UMTS and GSM. The inter-RAT handover is a hard handover.

C. Intra-PLMN Handover Versus Inter-PLMN Handover

Regardless of radio access system change or not, if the handover happens within the same network, i.e., having the same mobile country code-mobile network code, then it is named as intra-PLMN handover. And this includes intra-RAT handover and inter-RAT handover where the MCC-MNC is the same in both cases. Handover between different PLMNs, i.e., having different MCC-MNC, is named as inter-PLMN handover.

D. Intra-Cell Versus Inter-Cell Handover

The intra-cell handover is a handover within one sector or between different sectors of the same cell. Since this handover does not change cell, the related network connections will not be changed. The inter-cell handover is a handover between different cells, which requires related network connections to be changed accordingly.

E. Intra-Band Versus Inter-Band Handover

The intra-band handover and inter-band handover are handovers, where the frequency band of the target cell is the same as or different from that of the original cell. From the procedure point of view, parameters of frequency bands should be configured during regular measurement steps. Whether it is an intraband or an interband is transparent to other steps of the subsequent handover.

F. Cell Change Order and Redirection

The cell change order is a method that moves connected mode terminal from LTE access to GERAN. The LTE access network commands the terminal to move to the target GERAN cell with corresponding cell identity and frequency information. The terminal first goes to idle mode and then follows the access procedure of GERAN to get connected. During such a process, the user plane data transfer is interrupted. In order to accelerate the access process at GERAN and minimizes data transfer interruption, CCO is enhanced (that is network-assisted cell change, NACC) to convey target GERAN cell's system information to the terminal to avoid terminal reading system information when it accessing the GERAN cell.

The redirection is a method that moves connected mode terminal from LTE access to GERAN or UTRAN. The LTE access network releases the RRC connection while providing redirection information to the terminal. Such information includes frequency, cell identity, and system information of target cells. A terminal

camp at one target cell according to that information after RRC connection is released in LTE.

Both CCO and redirection use RAN Information Management (RIM) procedures to transfer information to and from source and target radio network via core network nodes of each system, e.g., MME and SGSN [15]. That avoids introducing new interfaces between different types of radio access networks.

Compared with handover procedure, which require negotiation about resources of terminal at target side between source and target systems, the CCO and redirection do not require such negotiation, the LTE system makes decisions by itself, based on necessary configured information.

5.4 Location Management from Core Network Aspect

5.4.1 *General Ideas of Location Management*

(1) Architectures and two basic functions

Location management from core network point view is consists of two functions, locationupdating and reachability management [15, 16]. Regarding the location updating, mobile terminal needs firstly registered with network once it powered up. After that, it is required to report to network both “I am alive” periodically, and “I am in a new location” when it moves into a new location area that it is not registered before. The core network entity that serves this terminal (i.e., received such reporting) remembers in which location area the mobile terminal is registered and refreshes mobile terminal’s “alive (i.e., reachable)” status once such reporting is received and terminal is allowed to access in such location. This serving entity also updated its identity to HSS.

For the reachability management, roughly speaking, when a mobile-terminated service targeted to a mobile terminal arrives, the network first finds out which core network entity currently the terminal is registered and then notifies this entity, then this entity will page within the registered location area of the idle-mode mobile terminal. Once the paging message is received, the mobile terminal contacts with the network immediately. By limiting paging messages within registered location area of the mobile terminal, paging within the whole network coverage is thus avoided at the cost incurred by location updating. With parameters tuning, e.g., size of location area and location-updating period, a reasonable trade-off between paging signaling cost and location-updating costs (includes signaling and terminal power consumption) could be achieved.

The function of location management described above was firstly implemented in 2G CS domain, and similar functions were also implemented in PS domain (both 3G and LTE), but with several enhancements, e.g., signaling optimizations within core network entities, combined location-updating procedures, and idle-mode signaling reduction. However, the basic location management architecture, that is,

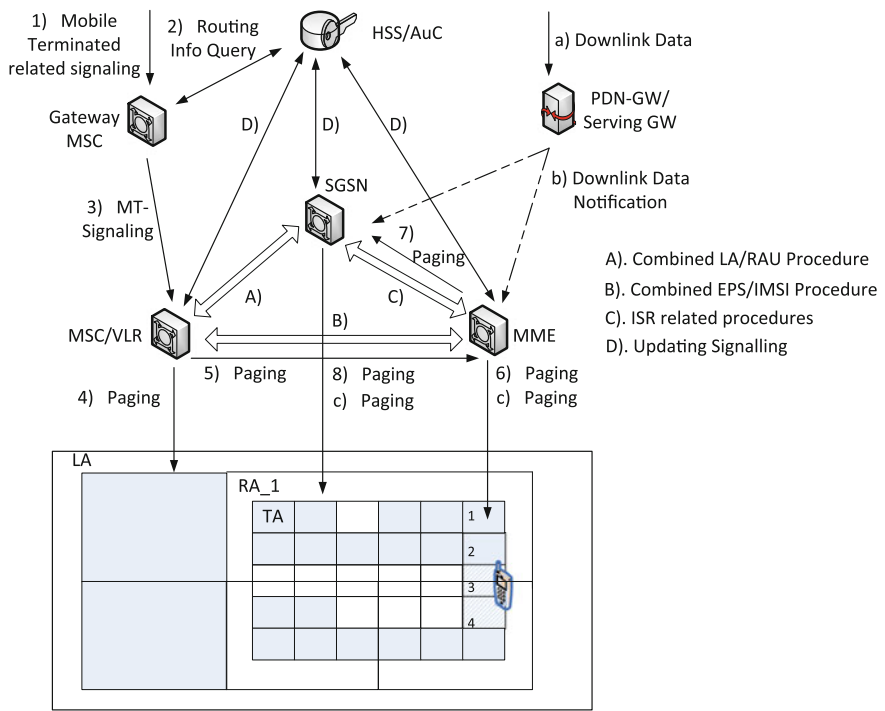


Fig. 5.18 Architecture, procedures, and LA/RA/TA of location management

two-level location management hierarchy, and those two basic functions, are kept in 2G/3G/4G systems.

Figure 5.18 summarizes several important aspects of location management of GSM/GPRS/LTE system, such as architecture hierarchy of location management, message flows regarding combined procedures and reachability management, and different concept of areas used in GSM/GPRS/LTE system.

(2) Location Area, Routing Area, and Tracking Area

The LA, RA, and TA are concepts used in location management at core network by mobility management entities, i.e., VLR of CS domain, SGSN of PS domain, and MME of LTE system, respectively. Table 5.6 gives out a comparison among LA, RA, and TA. Their identity (i.e., LA code, RA code, and TA identity) are broadcasted by their constituting cells, in each cells' system information block. By reading the SIB, the terminal knows whether it moved into a new area or not.

As an example, in the bottom part of Fig. 5.18, location areas (LA), routing areas (RA), and tracking areas are shown, with decreased sizes, where those sizes reflect most real network deployments, respectively. One or more RAs are always contained in one LA, and in most cases, the size of TA is smaller than that of RA.

Table 5.6 Comparison among LA, RA, and TA

	Description from comparison perspective
Location area	Within the CS domain, each location area consists of several GERAN/UTRAN cells. Once registered in a LA, terminal could move within this LA freely without updating its current location with the network, until periodical updating is needed. The VLR where terminal registered knows which LA the terminal is located, and VLR updates its identity to HLR if needed, e.g., at terminals first registration, or terminal moves and got registered at a new VLR
Routing area	Within the PS domain, each routing area is consists of several GERAN/UTRAN cells. The location management, such as registration and periodical updating procedures, is quite similar with that of CS domain. The major differences are that the size of an RA is smaller than that an LA, and it is the SGSN stores terminal's current routing area, and the SGSN updates its identity to HLR when necessary
Tracking area	Within the LTE network, each tracking area includes one or several E-UTRAN cells, and each E-UTRAN cell may belongs to more than one tracking area. However, the network knows the location of one idle-mode UE on the tracking area list granularity. The MME allocates a list that consisted of one or more TAs to terminal when it is registering with the network. The mobility pattern or history information of mobility of the terminal is considered when MME allocates such TA list to this terminal. Similarly, the MME also updates its identity to HSS when necessary.

Both LA and RA consist of many 2G/3G cells, and each 2G/3G cell is only belonging to one LA and one RA. However, one LTE cell could belong to more than one TA, and one TA could consist of one or more LTE cells. This gives more flexibility to design the size and shape of one TA and thus gives more flexibility to the location management of LTE.

Theoretically, the size and shape of one TA could be arbitrary, but it is carefully designed to meet special location management requirements of LTE network. In order to adapt to each terminals' mobility pattern, more than one TAs are organized as a set for location management purpose. By making the center of this set of tracking areas close to the terminal's current location, the chance of a UE rapidly making another tracking area update can be reduced. Thus, the size of one TA is smaller than that of one RA, in order to easily organize a good TA set for each terminal in a flexible manner.

(3) Location updating

During IMSI attach/GPRS attach or location area update/routing area update procedure, once the mobile terminal is accepted by the network, it is then registered in an LA or RA of a PLMN. The terminal simply knowing in which LA/RA it has registered by reading the LA/RA identities included in the accept message of corresponding procedure. The VLR and SGSN also record in which LA or RA the mobile terminal is located into a "context" of such mobile terminal, for future use. However, during EPS attach or tracking area update procedure, one or more tracking areas are allocated (denoted by a Tracking area ID list, TAI list) by MME

for this mobile terminal, as it is a registered area, and the TAI List is sent to the mobile terminal via attach and TAU accept messages. The MME takes mobile terminal's history mobility information such as last registered TAI into account when allocating such TAI List. By doing this, the shape and size of a registered area could adapt to mobile terminal's mobility pattern in EPS system, and this would reduce the amount tracking area update signaling.

During an attach procedure, the VLR, SGSN, and MME notify HSS with their identities. Due to user mobility, the current (old) serving VLR/SGSN/MME node may be relocated to a new node and the new serving VLR/SGSN/MME will notify HSS their identities during location-updating procedures, as well. The HSS records those identities received for future use, e.g., for reachability management propose.

During each procedure mentioned above, a timer value is also allocated by VLR/MME/SGSN and been sent to mobile terminal. Once returning to the idle mode, the mobile terminal starts a timer with this value and will perform periodical LAU/RAU/TAU procedure on the expiry of this timer (i.e., reporting "I am alive"). During idle-mode mobility, the mobile terminal contiguously reading camped cells location area identity (e.g., LA code or RA code or TA identity) and compares it with stored registered area's identity. If moved out from its registered area, the mobile terminal will perform a normal LAU/RAU/TAU procedure (i.e., reporting "I am in a new location").

For mobile terminals wish to use both CS service and PS services, they should be registered in both CS domain and PS domain. In order to minimize the signaling, separated registration procedures for CS and PS services are combined into one procedure, i.e., combined GPRS attach and combined LAU/RAU procedure, within GSM and GPRS system. In those procedures, the SGSN sends IMSI attach and location update request to MSC/VLR on behalf of mobile terminal via Gs interface, and then, a Gs association between SGSN and VLR for each mobile terminal is established and refreshed. For LTE system with CSFB architecture, combined EPSS/IMSI attach and combined track area update procedures are used to register mobile terminals in EPS system and CS domain, and a SGs association for each mobile terminal between MME and VLR is established and refreshed.

Detach procedure is used either by a mobile terminal to notify network it does not need CS and/or PS service any more or by network to notify mobile terminal that it could not get services from network. When explicit detach procedure (i.e., with explicit signaling need) is finished, the mobile terminal accurate location is not known by network since there is no way for network to contact with it. If network finds out that one mobile terminal did not contact with network for a quit long time, the network implicit detach the mobile terminal (i.e., no signaling between mobile terminal and network). In this case, it is also not possible to contact with the mobile terminal. Table 5.7 summarizes the above procedures.

(4) Reachability Management

Reachability management is a mechanism used to find a mobile terminal in idle mode, when the mobile-terminated service is arrived for such mobile terminals. As

Table 5.7 Procedures for location registration in GSM/UMTS/LTE

	CS domain	PS domain	LTE
Initial registration procedures	Location area update	GPRS attach procedure (for only PS services); combined GPRS attach procedure (for both PS and CS services)	EPS Attach Procedure (for only PS services); combined EPS/IMSI attach procedure (for both PS and CS services)
Location update procedures	Location area update	Routing area update procedure; combined routing area/location area update procedure	Tracking area update procedure; combined tracking area/IMSI update procedure
Deregistration procedures	IMSI detach (mobile terminal or network initiated)	GPRS or IMSI Detach (mobile terminal or network initiated)	EPS or IMSI detach (mobile terminal or network initiated)

shown in Fig. 5.18, for CS services, such as mobile-terminated CS voice calls, reachability management includes following steps: (1) the MT voice call setup signaling arrives at Gateway MSC, (2) G-MSC query HSS with called party’s MSISDN to retrieval current serving MSC/VLR of called MS, (3) G-MSC route the call step signaling to the serving MSC/VLR, and (4) the serving MSC/VLR sends paging messages with “CS domain” indication within the called mobile terminals registered location area, and those paging messages will be broadcasted in all 2G/3G cells belongs to this location area. Once received the paging messages, the mobile terminal establishes an AS layer connection (i.e., RR connection) and indicated to NAS layer once succeeded. Then, the mobile terminal waits for more command from network, and mobility procedure may be imitated by network or terminals. (5) If there is a SGs association for this mobile terminal (i.e., the CSFB is deployed and this mobile is EPS/IMSI attached), the MSC/VLR does not page mobile terminal in location area, instead it sends a SGs Paging message to MME, (6) the MME sends paging messages with “CS domain” indication within the called mobile terminal’s registered TAIs, and those paging messages will be broadcasted in all 4G cells belonging to those registered tracking areas. If Idle-mode Signaling Reduction has been activated (this would be discussed in Sect. 5.4.2), then MME also sends paging to SGSN in step (7) to trigger SGSN sending paging messages within this terminal’s registered routing area. Once received paging message with “CS domain” indication in LTE cell, the mobile terminal responses by initiating extended service request procedure to the MME. Then, MME will perform a handover procedure to move this terminal to MSC/VLR to answer call setup signaling.

For PS service, the procedure is quite different. For PS services, a PDP context must be activated in GPRS network or a PDN connection must be established in EPS network, before the mobile terminal can communicate with external packet data networks. In order to support “always-on” feature, a PDN connection to an external packet network (e.g., Internet or IMS domain, controlled by operator via

subscription method) must be established successfully during the EPS attach procedure; otherwise, the attach fails. However, the PDP context within GPRS system is activated after a successful GPRS attach procedure. Once activated/established, the PDP context and PDN connection are maintained during mobility procedure, until been deactivated or torn down. The GGSN/PDN-GW is a top-level user plane anchor for mobile terminal and will not change during mobility procedure. However, the second-level anchor SGSN and Serving GW may be relocated during mobility, which are mainly based on the user's current location and network topology. If relocated, the GTP tunnel of established PDN connections between old second-level and top-level anchor would be moved to between the new second-level anchor and top-level anchor. Thus, the GGSN/PDN GW always knows the identity of a second-level anchor.

As shown in Fig. 5.18a, when the downlink data targeted to a mobile terminal arrive at GGSN/PDN-GW, they will be encapsulated in GTP-U packets and then send to SGSN/Serving GW on the corresponding GTP tunnel of this terminal. Once SGSN/Serving GW receives the first downlink GTP-U packet, if there is no GTP-U tunnel for this packet between SGSN/Serving GW and RAN nodes, paging will be initiated as follows. In GPRS system, since the SGSN knows in which routing area the mobile terminal is located, it pages the mobile terminal within the routing area, i.e., the step 8) in the figure, sends Iu-PS or Gb interface paging messages to every RAN controllers (RNC or BSC) who have cells serving this routing area. In EPS system, the Serving GW buffers the received GTP-U packets and then sends a downlink data notification to MME as in step b). The MME finds out the target mobile terminal's registered TA list, and pages in those tracking areas belong to the list, i.e., the step 6) in the figure, sends S1 interface paging messages to each eNodeB who have cells within a tracking area contained in the List. Once received paging messages over the air, the mobile terminal makes response with network, by initiating the service request procedure. When such a procedure completed, the radio bearers over the air and GTP-U tunnels over the Iu/Gb/S1 interfaces have been established, and the data communication between mobile terminal and packet data network begins.

(5) Paging Repeat and Priorities

The radio propagation environment is very complicated and unpredictable; and the paging message may not reach the targeted mobile terminal, e.g., the terminal may located at street corner, within elevator and other places where radio coverage is not good, for a very short period time due to user mobility. In such case, resending the paging message would increase the possibility that the mobile terminal responses to paging. Based on such consideration, paging is repeated several times decided by configuration. The MSC/VLR, SGSN, and MME are network nodes that issue paging messages to RAN and control how paging is repeated, e.g., the interval of sending paging messages, and how many times the paging messages should be sent until answered or stop sending. For priority services, corresponding paging messages may be treated in a priority way by SGSN and MME.

The RAN node pages mobile terminals over air on specific timeslots of one or more radio frames. The RAN node and terminal use same algorithm and the same input (e.g., mobile terminal's temporary identity) to figure out exactly which timeslots to send and to receive paging messages.

5.4.2 *Idle-Mode Signaling Reduction*

Multi-RATs supports of UE and network are common due to network evolution and competition reasons. Hence, the idle-mode mobility between LTE and UTRAN/GERAN is common, especially for a spotty coverage case, where the LTE will be first deployed as a kind of "hot spot" to offload data traffic. How to reduce the signaling between the UE and network as well as signaling within the network is one important problem that must be resolved for this case [15, 18].

The idle-mode signaling reduction (ISR) feature is the answer. The basic idea is that the UE is registered at both an UTRAN/GERAN Routing Area (RA) and an E-UTRAN Tracking Area (TA, or multiple TAs, identified by a TA list). The network and the UE both keep the two registrations in parallel; and there is no updating signaling needed when the idle-mode UE moves from the registered RA to the registered TAs, and vice versa [18].

Such kind of dual registration is achieved by "ISR activation" operation. The UE first attached to network via one RAT, e.g., UTRAN, and got registered at SGSN and HSS. When reselected to LTE, UE initiates TAU procedures rather than EPS attach procedures. During TAU procedure, the MME and SGSN negotiate their ISR support capability, if allowed, the MME indicate ISR would be activated to SGSN and Serving GW, and then indicates "activate ISR" in TAU accept message sent to UE. During the TAU procedure, the MME also updates HSS with its identity. By doing this, this UE is then "dual-registered" at SGW and HSS (i.e., the HSS stores both serving MME and serving SGSN address for this UE, and Serving GW also has control plane connection with both MME and SGSN, for this UE). The UE also keeps two contexts for UTRAN and LTE, e.g., temporary identities, P-TMSI and GUTI, UMTS and EPS security contexts, and bearers/PDP contexts.

If there is downlink data comes from the PGW to Serving GW for a UE, the Serving GW requests both MME and the SGSN to page UE, if the Serving GW knows ISR is activated for this UE. Both MME and SGSN page the UE within its registered TAs and RA, respectively. Once received the response via one RAT, e.g., the E-UTRAN, Serving GW stops paging on the other RAT. i.e., Serving GW requests the SGSN to stop paging.

During every RAU and TAU procedure, the network will refresh the ISR status, once it has been activated. If the ISR active status could not be maintained by network, the SGSN or MME deactivates ISR by not including "activate ISR" indication in RAU/TAU accept messages. The UE may also locally deactivates ISR without notifying network in some scenarios. The main reasons behind "ISR deactivation" operation include the following: (1) some network nodes may not

support ISR, (2) modified bearer/PDP context, or changed DRX settings (which changes paging occasions) cannot be synchronized by using existing procedure, and (3) dual registrations cannot be maintained at Serving GW. Detail scenarios that ISR should be deactivated could be found in TS 23.401 [15].

5.4.3 *Typical Location Management Procedures*

In this section, EPS attach, tracking area update, and detach procedure are described [15].

(1) EPS Attach Procedure

The attach procedure is initiated by a terminal to register itself to network. For terminal initially attached to PS domain of GPRS network, no connections to external packet network are established, i.e., no PDP context would be activated. While during attach to LTE network, a PDN connection is established. For those attach procedures, the terminal must pass authentication before get authorized for using any services. During GPRS attach or EPS attach procedure, the SGSN or MME may register this terminal to CS domain on behalf of this terminal if terminal requests so in the attach request message. An example of such procedure, named as combined EPS/IMSI attach [15], is shown in Fig. 5.19.

The attach request (in step 1) is a non-access stratum message, the eNodeB does not interpret any NAS message, but it knows that this message should be send to an MME. The eNodeB selects one MME according to network topology and loads balancing consideration if no native GUMMEI is found in the RRC message carrying the NAS messages. The GUMMEI is short for Globally Unique MME Identifier, which is a part contained in the Globally Unique Temporary Identity (GUTI). The GUTI is allocated by MME during EPS attach, GUTI reallocation, and possibly during tracking area update procedure. The terminal stores this GUTI even if it is detached. The stored GUTI would be signaled to network during next EPS attach or GPRS attach procedure (in this case, the GUTI is mapped to P-TMSI by terminal).

Once received attach request with GUTI, the new MME first uses this GUTI to retrieve mobility management context locally (i.e., within itself memory or storage), if not found, then MME retrieves context from the old MME identified by GUMMEI via identification request/response procedure (step 3). If still not found, the MME requests the terminal to send user's IMSI (step 4), and a full authentication and security control procedure are performed (step 5), where mutual authentication between terminal and network is performed, and ciphering and integrated protection for subsequent signaling may be activated. If the attach request is received without integrated protection, the step 5 is then mandatory. Depending on the operator's policy, MME may be configure to perform authentication procedure even MM context could be found locally or from old MME.

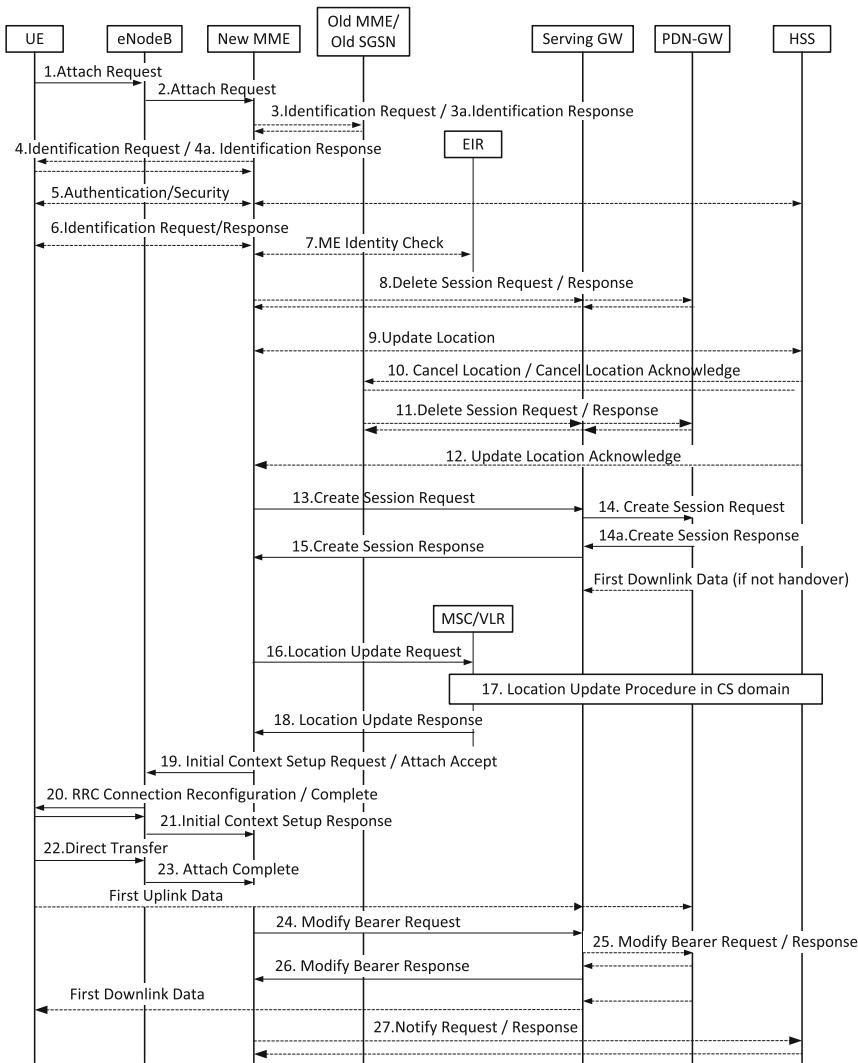


Fig. 5.19 EPS initial attach procedure. Reprinted from 3GPP TS23.401 [15]

Depending on operator's policy, MME may retrieve terminal's identity (i.e., IMEI) and also its software version via a security protected Identification Request/Response (step 6), and MME checks with EIR whether this terminal is prohibited (e.g., a stolen terminal) to enter the network or not (step 7).

If there are active bearers for this terminal (in some case, the terminal reattaches to the same MME without having properly detached before), the MME deletes these bearers (step 8).

If the MME is changed, i.e., the GUTI indicated by the terminal was not allocated by the new MME, and then this MME notifies HSS its identity (i.e., to update serving MME identity of this terminal stored in HSS) (step 9). If no valid subscription data in this MME for this terminal, this step is also used to retrieval subscription data from HSS. In this case, the HSS should notify (cancel) the terminal previously registered MME/SGSN since this terminal has registered in the new MME, then the old MME/SGSN deletes terminal's MM context, and bearer-related resources, if any, by using delete session request (step 10 and 11).

During the default PDN connection establishment procedure (from step 13 to 15), the PDN GW allocates IP address (IPv4 or IPv6 or both, depends on PDN type negotiated between terminal and network) for this terminal. Which PDN is default is stored as subscription information; however, the terminal may use ciphered option request/response procedure to notify network to establish PDN connection other than the default one.

The MME registers this terminal with CS domain if has been requested by the terminal. The MME first figures out a MSC/VLR on local configuration and then sends location update request to it (step 16). Then, this MSC/VLR performs location update for this terminal and returns result to the MME (step 18).

Then, the MME triggers radio bearer setup by sending initial context request to eNodeB, which is a S1-AP message, with the NAS message attach accept piggy-backed. The allocated IP address, GUTI, TAI list, and result of CS domain registration are included in attach accept message (step 19, 20, 21, 23). Once the RRC connection is ready, the terminal sends an attach complete message to network to acknowledge the received GUTI and TAI list (step 23). And the terminal could send uplink data by using allocated IP address.

Upon the receipt of both initial context response and attach complete message, the MME updates established bearers information (e.g., eNodeB addresses and tunnel endpoint identifier for downlink) with Serving GW (and PDN GW may be involved) according to result of the radio bearer establishment (step 14). Then, Serving GW starts to forward downlink data to eNodeB.

For terminal support handover to non-3GPP radio access, e.g., Wi-Fi, the MME check that whether the PDN GW selected is same as that received in the subscription data. If not, the MME sends this selected PDN GW's identity to HSS (step 27). When the terminal later handovers to non-3GPP radio access, the same PDN GW is used, and the same IP address is allocated to this terminal. If the terminal first attaches via non-3GPP radio access and then needs to use 3GPP radio access later, such as LTE, the terminal indicates "handover attach" in the attach request. Once received this indication, the PDN GW identity is retrieved from terminal's subscription by MME, and a PDN connection would be established to the PDN GW identified by this identity.

(2) Tracking Area Update Procedure

The tracking area update procedure is used by the terminal to update its "I am alive" state periodically and notify network that it has moved out of its registered

tracking areas (those tracking areas listed in the TAI list is allocated to the terminal during attach, TAU, and GUTI reallocation procedure). Tracking area update procedure may also be used for other purposes, such as load balancing (follows a S1-release procedure which a “load balancing tau required” indication sent to terminal), updating changed UE network capability, and attaching to CS domain [15]. In Fig. 5.20, the combined tracking area update and location area update procedure, where MME and Serving GW are both relocated, are shown as an example.

The eNodeB first checks whether it has connection with the MME identified by the GUMMEI in the received RRC message that carries the TAU requests (step 1). If not, the eNodeB sends this message to an MME selected based on network topology (step 2). Since this MME is not the MME who allocates the GUTI included in the TAU Request, the new MME queries the old MME for this terminal’s context (both mobility context and bearer context) with the received GUTI (step 3).

If the integrated protection check of the TAU request fails, authentication and security procedure is mandatory (step 4). If the MME figures out that Serving GW should be relocated, e.g., based on network topology, a new Serving GW is expected to serve the UE longer and/or with a more optimal UE to PDN GW path, or if a new Serving GW can be co-located with the PDN GW, then new MME indicate “serving gateway change” to old MME/old SGSN in context acknowledge (step 5).

The old MME/old SGSN knows that new MME has relocated Serving GW, and then it marks terminal’s bearers context as invalid (step 5). If terminal come back to old MME/old SGSN by imitating TAU or RAU procedure before the completion of this ongoing tracking area procedure, the old MME or old SGSN first deletes session on old Serving GW and PDN GW and re-creates sessions on old Serving GW and PDN GW. Such a design is to handle “PING-PONG” movement scenario.

The new MME compares the bearer status received from terminal and those bearer status contained in context of this terminal received from old MME or old SGSN. The MME releases any resources related to those bearers that are not active in the UE. If there is no bearer context at all, the MME rejects the TAU Request. How could no bearer context happen for a terminal? Here is the answer. If a terminal has registered via 2G/3G and reselected to an LTE cell, it imitates TAU procedure to fulfill its registration at EPC network. In this case, the P-TMSI and P-TMSI Signature allocated by SGSN is mapped to a mapped GUTI by terminal. The terminal provides this mapped GUTI during this procedure, and then, MME could find the correct SGSN to fetch terminal’s context. In this case, the terminal may have no activated PDP context at 2G/3G side; thus, no bearer context exists for this terminal.

For those bearers with active status, the new MME triggers bearer establishment at target side (step 6). Then, Serving GW address and tunnel endpoint identifiers for uplink traffic for those bearers are returned to new MME (step 8).

Then, the new MME notifies HSS that the terminal has moved to this new MME (step 9), and HSS stores the identity of this MME for this terminal and then send

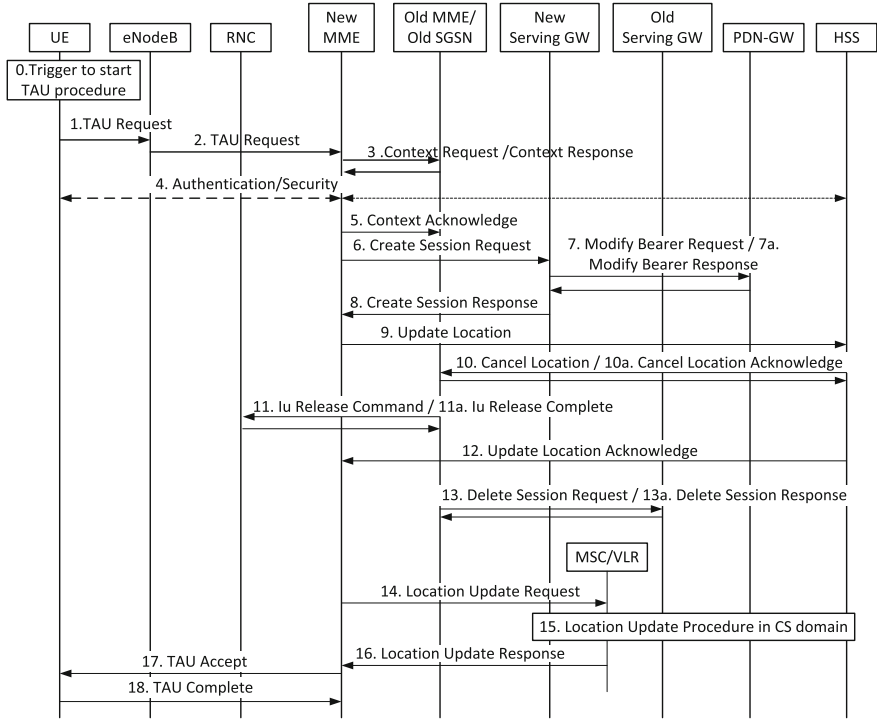


Fig. 5.20 Tracking area procedure. Reprinted from 3GPP TS23.401 [15]

cancel location to the old MME or old SGSN (step 10), to let this terminal’s bearer-related resource to be released at source side (step 11, 11a).

If the terminal has requested to register itself in CS domain, i.e., to perform IMSI attach, or to update it status at CS domain, the MME sends location update to an MSC/VLR, to trigger the location update procedure within CS domain (step 15). The result returned by MSC/VLR (step 16) is then returned to terminal in TAU accept (step 17).

If the terminal does not require “enable data sending” in the TAU request, or there is no downlink data pending during the TAU procedure, the MME release signaling connection after receiving the attach complete. Otherwise, the sending of TAU accept is in conjunction with the user plane step procedure, i.e., the MME triggers eNodeB to setup radio bearers and context for this terminal.

(3) Detach Procedure

The detach procedure is used to deregister a terminal from network. This procedure can be initiated by either terminal or network. Figure 5.21 shows an example, which is a UE initiates detach procedure when ISR has been activated. The terminal sends detach request to SGSN, and SGSN releases bearers for this terminal by sending delete session request to the Serving GW (step 2).

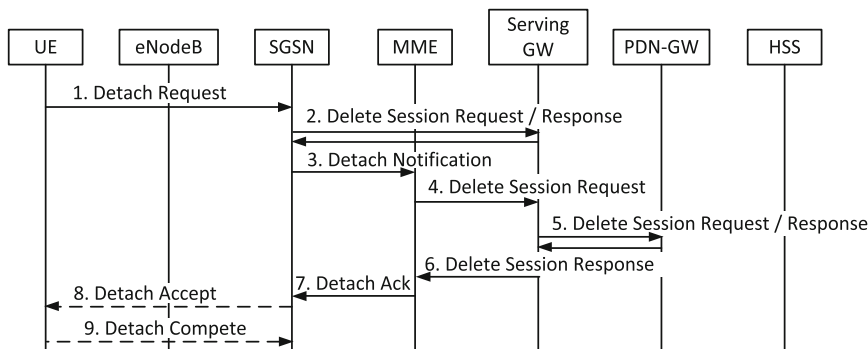


Fig. 5.21 Detach procedure. Reprinted from 3GPP TS23.401 [15]

Since knowing that ISR has been activated for this terminal, the SGSN notifies MME where this terminal is registered that this terminal is detaching (step 3). Then, MME releases bearers at LTE side (begins with step 4). Knowing ISR is activated, the Serving GW only notifies PDN-GW (step 5) after receiving delete session request message from both side. Upon receiving response from PDN-GW, the Serving GW returns response to MME (step 7), and then MME returns detach acknowledgment to SGSN.

Then SGSN sends detach accept to terminal, if signaling connection still exists to this terminal (since terminal may switch off immediately after sending detach request).

The network may also detach terminals, with or without detaching-related signaling between network and terminal, i.e., explicitly or implicitly detach, respectively, in following cases:

- Due to operator-determined purposes, HSS sends cancel location request to MME or SGSN to deregister a subscriber from the system;
- Once found out that the terminal has not contact with network upon the expiry of implicit detach timer, MME/SGSN implicitly detaches this terminal, or due to OAM intervention, the MME/SGSN explicitly detaches the terminal by sending a detach request to the terminal;
- Once determined that GW relocation is desirable for all PDN connection(s) serving SIPTO-allowed APNs, the MME explicit detaches the terminal by sending a detach request to the terminal, with “reattach required” indication included. This only happens for the idle-mode terminal.

During detach procedure, resources related to PDN connection would be released; however, mobility-related information (such as temporary identity, subscription information, and keys derived for security protections) of this terminal is kept by MME and SGSN for future use after completion of this procedure for a configured time period (could be up to 48 h), in order to save signaling (especially signaling to the

HSS) when this terminal reattached to network. Once mobility-related information is deleted (after such time period ends, or be configured to be deleted immediately after detach), HSS would be notified of such deletion.

5.5 Handover Control from Core Network Aspects

Broadly speaking, mechanisms that ensure ongoing service not interrupted (or, such interruption is not noticed by user) during user mobility are considered as handover control. At radio network layer, handover control moves user service's radio channels from source/old cell to target/new cell. If radio network nodes that control the source and target cell are connected to different core network nodes, then core network nodes are involved. From core network aspects, handover control establishes transport network layer connection between anchor core network node and target core/radio network nodes and moves user service from old transport network layer connection to new connection. Increasing handover success rates, minimizing handover delay, avoiding packet dropping/service downgrading, reducing control-signaling costs are goals of handover control must achieve. Handover procedures for both CS service and PS service are detailed in this section, from core network point of view.

5.5.1 *Handover Control Within CS Domain*

In Fig. 5.22, an inter-MSC handover procedure with GERAN is shown, which happens within CS domain. When a mobile terminal initiates a voice call, say to a POTS terminal, the MSC/VLR that serves this mobile terminal takes full control of this call, e.g., call setup, charging aspect and handover control. This MSC is referred anchor MSC, and this anchor MSC would not change due to user mobility and remains unchanged till this call finished.

Once the source BSC decided that a handover is needed, it sends handover required message to MSC (step 1), with target cell identity included. If the MSC found out that target cell is controlled by another MSC (e.g., MSC_B in this case), a perform handover message is been sent to this target MSC (step 2). The target MSC prepares radio channel (step 3, 4), and sets up connection for speech between anchor MSC and itself (step 6). Once succeed the anchor MSC issue a handover command to source BSC and mobile terminal (step 7, 8). The mobile terminal would handover to target cell after receiving such command. Once the handover is detected by target BSC, the target BSC notifies target MSC, and target MSC then notifies the anchor MSC (step 10, 11). The anchor MSC then switches voice stream to target MSC and releases radio and network resource related to this call at source side.

During the procedure, a handover number (HON) is allocated by target MSC for this mobile terminal. The HON is used by source MSC to route call-related

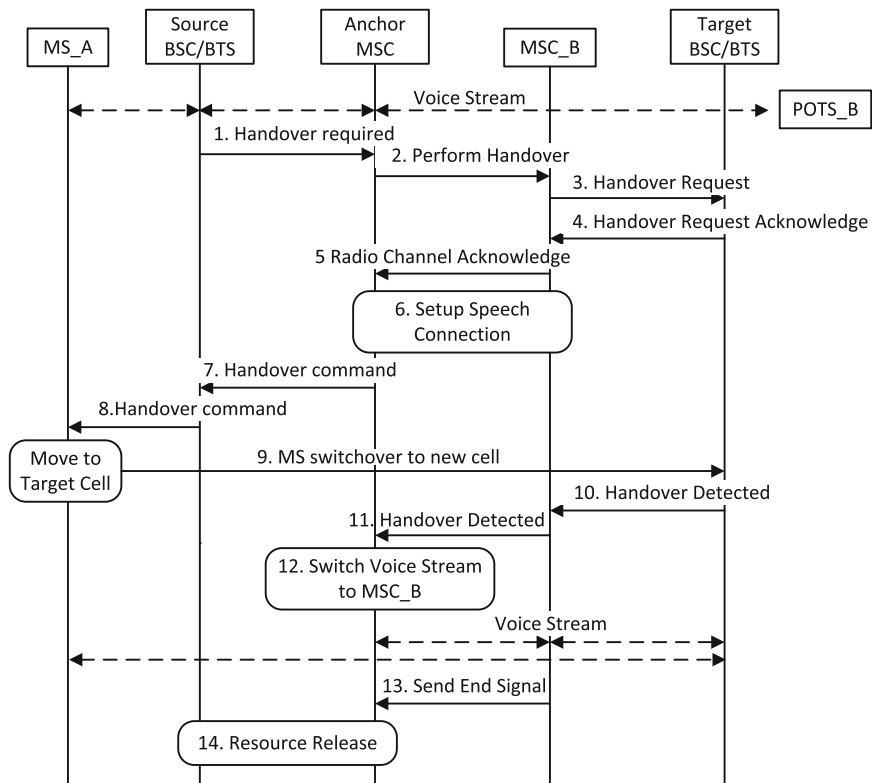


Fig. 5.22 Inter-MSC handover procedure

signaling and establishes speech connections between MSCs. After this basic inter-MSC handover, if this mobile terminal with one ongoing voice calls is going to move into another cell controlled by to a third MSC, say MSC_C, a subsequent inter-MSC handover would be performed. In such case, the BSC sends a handover required message to MSC_B, with target cell identity. The MSC_B figures out that target cell belongs to MSC_C, and then MSC_B send the perform handover message to the anchor MSC. The anchor MSC will contact MSC_C for fulfill this further handover procedure.

5.5.2 Handover Control Within PS Domain

(1) Overview

In order to maintain IP network layer (and above) service continuity, mobile terminal’s IP addresses used to communicate with external packet data network are

kept unchanged during any mobility procedure. Despite allocating IP addresses to the terminal or not, the PDN GW or GGSN is always used as user plane anchor for the mobile terminal's IP services, and thus, it will not be changed/relocated to another one if there are ongoing terminal's IP sessions.

User IP packets that to and from external PDN and terminal are transferred by tunnels within PS domain and bearers within radio access network where the terminal is current connected. The handover control within PS domain is focused on moving those tunnels and bearers to accommodate user mobility, while at the same time without interrupting, delaying ongoing user packets transfer, or degrading QoS committed for such user packet transfer, or dropping user packets that should not be dropped. For both GPRS and LTE network, the "make-before-break" approach is adopted, i.e., network resources at the target side are reserved prior to commanding terminal to move to the target cell. Mechanisms used to establish tunnels and reserve bearers resources at the target side, to switch packets forwarding from old tunnels/bearers to new one, to release old tunnels/resources at the source side are carefully designed to ensure the success of handover. During handover, radio network nodes (e.g., BSC or RNC or eNodeB) and core network nodes (e.g., SGSN or SGW or MME) may be relocated due to old nodes could not serve the terminal. In some cases, in order to optimize packet transfer route, i.e., to reduce intermediate nodes of layering IP tunnels (or shorten tunnels between PGW/GGSN and radio network nodes from transporting network layer point of view), SGW/SGSN may be relocated even if the old one still could serve for the terminal. For establishing/releasing/modifying/moving such tunnels, the famous GTP protocols are widely used within most operators' PS domain (both GPRS and LTE), while some operators choose the PMIP as an alternative in LTE network. The moving/mobility of bearers is transparent to user IP packets transfer between the terminal and external IP network via the PDN GW/GGSN.

(2) PS Handover

Packet-switched (PS) handover is firstly introduced for GERAN, which is used to handover a mobile terminal with one or more packet flows from a source cell to a target cell, where at least one of the cells is a GERAN cell [42]. During this procedure, real-time packet switching with strict QoS requirements on low latency and packet loss are supported via the "make-before-break" approach.

Depending on the locations of source and target cell, PS handovers are classified into intra-BSS HO (located within the same BSS), intra-SGSN HO (within different BSSs and belong to the same SGSN), and inter-SGSN HO (within different BSSs and belong to different SGSN). If source and target cell use different radio access types or use different type of interface connect to core network (i.e., different modes), then this is an inter-RAT/mode HO. All those HO procedure consists of a preparation phase and an execution phase.

One typical procedure of inter-RAT/mode and inter-SGSN handover, GERAN A/Gb-mode to UTRAN/GERAN Iu-mode handover, is shown in Fig. 5.23.

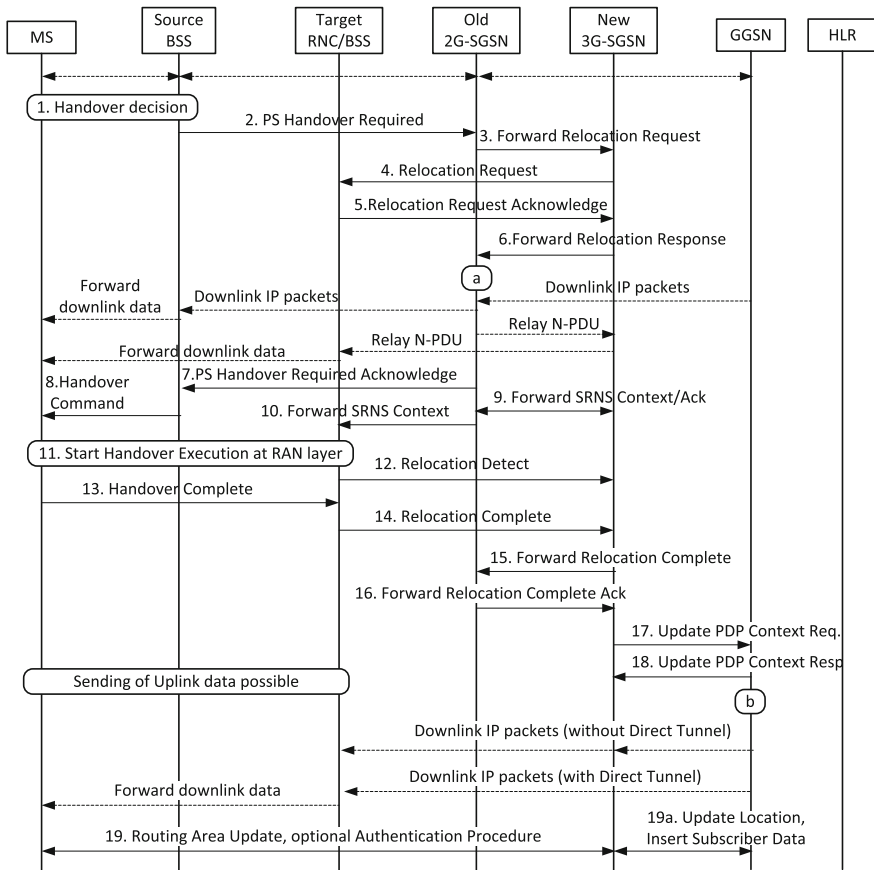


Fig. 5.23 Inter-RAT/mode handover procedure, from GERAN to UTRAN. Reprinted from 3GPP TS23.401 [15]

During preparation phase, source BSS decides and notifies 2G-SGSN a PS handover is required with Target RNC Identifier, source cell identity, a Source RNC to target RNC transparent container, active packet flow Context lists (which describes packet flows that needs resources allocated during preparation phase) (step 1, 2). Based on the target RNC identifier, 2G-SGSN figures out that this is an inter-SGSN handover and further determines the target 3G-SGSN and sends forward relocation request to this 3G-SGSN with MS context (both mobility-related and PDP-related parameters) included (step 3).

The 3G-SGSN reserves resource at target side (e.g., send relocation request to target RNC/BSS, which includes either GGSN address and tunnel endpoint identifiers for uplink data if direct tunnel feature is enabled or 3G SGSN address and tunnel endpoint identifiers for uplink data, and then receive Relocation Acknowledgment from target RNC/BSS, which contain radio access bearer setup

list and RAB failed to be setup list; for each RAB to be setup, it is defined by the target RNC/BSS address and the downlink tunnel endpoint identifier for user data), then return forward relocation response with following parameters, list of set up pfcS, SGSN addresses, and tunnel endpoint identifiers for both control plane and user plane (for , those addresses and identifiers are used by 2G SGSN to forward data during the handover procedure), target RNC to Source RNC container, to old 2G SGSN (step 6). Once received the response, 2G-SGSN considers preparation phase is finished (point a), and acknowledges source BSS, to let it command mobile terminal move to target cell (step 7).

Within the handover command sent to terminal, target cell-related information is included, such as radio resources allocated in target cell, system information needed for access the target cell (step 8). At the same time, 2G-SGSN duplicates downlink packets sent to source BSS and send them to 3G-SGSN. The 3G-SGSN sends those packets to target RNC/BSS, and RNC/BSS may start blind transmission to terminal on those allocated resources. For those packet flow requires ordered delivery, old 2G SGSN notifies 3G SGSN the sequence numbers of the GTP-PDU next to be transmitted in the uplink and downlink direction in Forward SRNS context message (step 9). After the forward SRNS context message is sent to RNC/BSC, further uplink N-PDUs received by the old 2G SGSN from the source BSS would not be forwarded to the GGSN.

After knowing that terminal has completed the handover to target cell by receiving handover complete message, the RNC/BSC indicates (step 14) the completion of the relocation, and then, 3G SGSN is ready to receive uplink packets from target RNC/BSC. If received any uplink packets from RNC/BSS, the 3G SGSN forwards those packets direct to GGSN. After notifying forward location complete to old 2G SGSN (step 15), the old 2G SGSN starts a timer after sending back acknowledgment. Upon the expiry of this timer, the old 2G SGSN stops forwarding packets to the new 3G SGSN. The new 3G SGSN sends an update PDP context request to GGSN after receiving acknowledgment from 2G SGSN. If direct tunnel is used during preparation phase, RNC/BSS addresses and tunnel endpoint identifiers and a DT flag would be provided to GGSN, instead of the new 3G SGSN addresses and tunnel endpoint identifiers, in this request message (step 17). After response this request, the GGSN forwards downlink packets to either RNC/BSS if direct tunnel is used or 3G SGSN, instead of 2G SGSN.

If has been handover to a new routing area, the terminal should initiate the routing area update procedure (step 19).

By using direct tunnels, the packet transfer delay is reduced, packet forwarding from old 2G SGSN to new 3G SGSN and ordered delivery function avoids packet dropping and ensuring in-sequence packets delivery during the handover procedure.

(3) SRNS Relocation

Within 3G networks, an interface named as Iu-r is introduced between two RNCs to reduce delay for both CS handover and PS handover, by not involving core network entities during the handover procedure. Instead, source RNC mimics

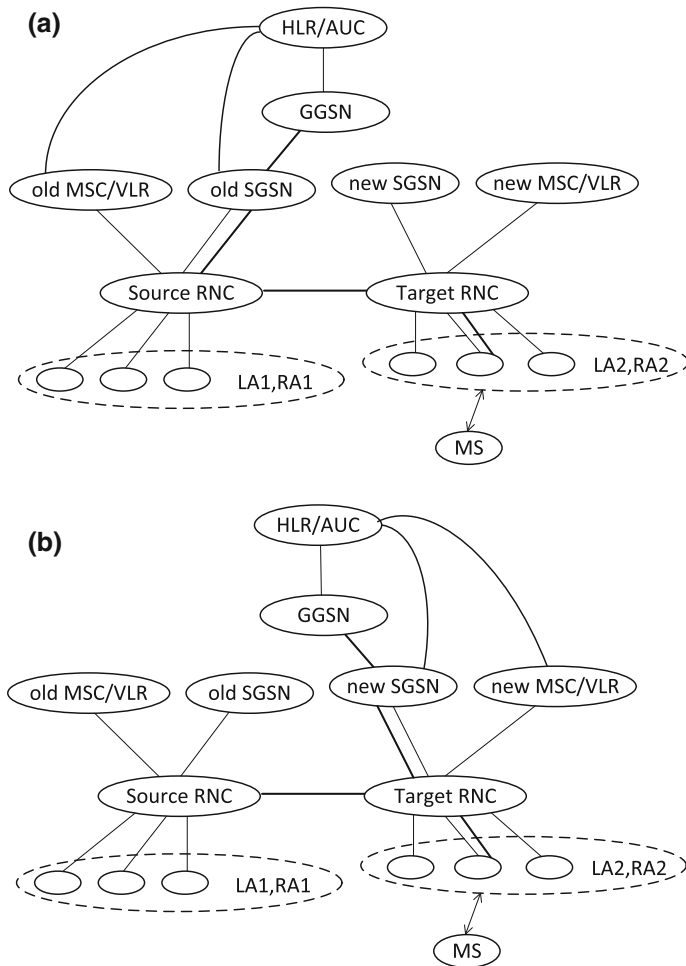


Fig. 5.24 Snapshot of network with SRNS Relocation procedure. **a** Before and **b** After. Reprinted from 3GPP TS23.060 [16]

the MSC or SGSN to control the handover, e.g., reserve resources at target RAN, and command terminal to move to target cell.

After such handover, in order to reduce transmission delay and resources usage, SRNS relocation is designed to move the RAN to CN connection point at the RAN side from the source SRNC to the target RNC, from a “standing still position,” i.e., Iu links are relocated, please refer to Fig. 5.24.

If the target RNC is connected to the same SGSN as the source SRNC, then it is an intra-SGSN SRNS relocation; further, an intra-SGSN routing area update is followed if the terminal moves to a new routing area.

(4) S1 Handover

The S1-based handover is used when the X2-based handover cannot be used, e.g., the target eNodeB has no connection with the source eNodeB, or the source eNodeB knows that the X2 handover is not possible or allowed based on local configuration, or after a failed X2 handover to the target eNodeB and etc.

It is source eNodeB who decide imminent handover is an X2 handover or an S1 handover based on above information upon receiving the terminal's measurement report (step 0). Please refer to Fig. 5.25 for details. The S1 handover procedure on LTE side is also used when the source access system is GERAN or UTRAN and the target access system is LTE.

The source MME based on the received target cell-related information in step 1, such as target eNodeB identity, target tracking area identity, to decide whether to select a new MME. The source MME selects a target MME that serves the target tracking area, if the source MME could not serve this terminal or the target tracking area (step 2).

When the target MME finds out that the source Serving GW cannot continue to serve the terminal, it would select a target Serving GW based on network topology, which could reduce the probability of changing Serving GW in the future. The source MME may also select a new Serving GW if the source MME continues to serve this terminal.

If a new Serving GW is relocated, resources for tunnels between the target eNodeB and the target Serving GW are reserved (step 3), e.g., the Serving GW addresses and tunnel endpoint identities for each bearers to be used for the uplink traffics are returned to MME (step 3a).

The MME requires target eNodeB to prepare bearer resource (step 4). The Serving GW address and tunnel endpoint identity for each bearer to be set up, and corresponding QoS parameters are indicated to target eNodeB. The target eNodeB returns which bearers have been set up successfully and which have not to MME (step 4a). For those bearers have been setup, the eNodeB address and tunnel endpoint identity used for downlink traffics for each bearers are also signaled.

There is a small time period between the source eNodeB sends handover command (step 9) to terminal and PDN GW switches downlink data transfer to target Serving GW(right after step 18a). During such period, downlink data are still sent to source eNodeB. For those bearers (decided by eNodeB) that carries user data requiring lossless and ordered delivery, data forwarding of downlink data is performed.

Direct forwarding tunnels between source and target eNodeB should be established and used if it is possible. Otherwise, the source eNodeB indicates this to MME in step 1, and indirect forwarding should be used instead. If MME decided to use indirect forwarding, then tunnels for indirect forwarding are established from source eNodeB to target eNodeB, via source Serving GW and via target Serving GW if Serving GW is relocated. The node addresses and tunnel identifiers of such tunnels are signaled in step 4a, 5, 5a, 6, 7, and 7a. The indirect forwarding tunnel for downlink data, with and without Serving GW relocated, is shown in Fig. 5.26.

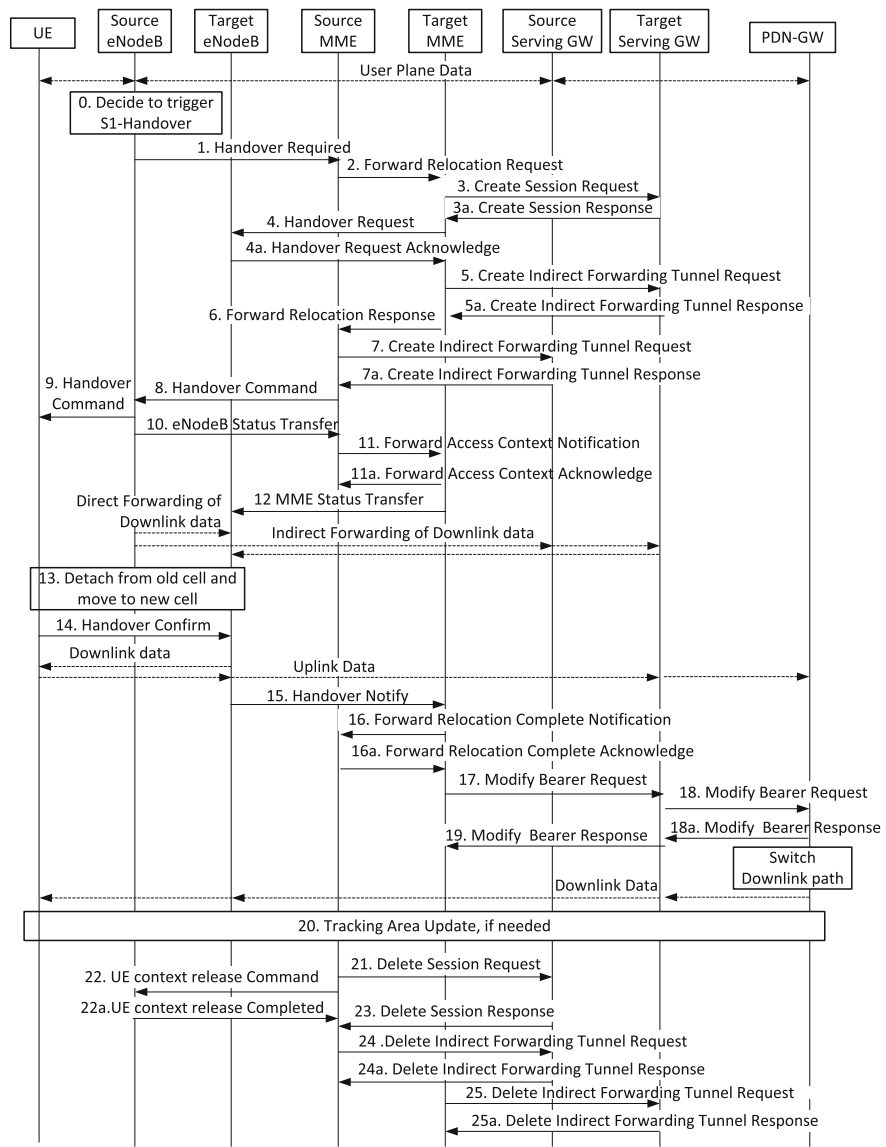


Fig. 5.25 S1 handover, with MME change and/or Serving GW change. Reprinted from 3GPP TS23.401 [15]

To fulfill lossless and in-order delivery (i.e., PDCP status preservation) for user data subject to indirect forwarding, the source eNodeB sends an eNodeB status transfer-related information such as PDCP and HFN status of those radio access bearer to facilitate PDCP status preservation (step 10 to 12).

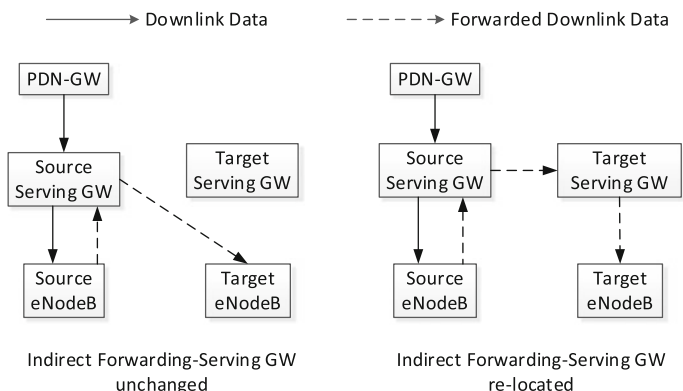


Fig. 5.26 S1 handover, showing indirect forwarding path for downlink data

After confirming that a terminal has moved to target side, the target eNodeB starts forwarding received forwarded data from the source side and notifies MME that the terminal has completed the handover to the target eNodeB. Upon receipt of such confirmation (step 16), the source MME (regardless having been relocated or not) starts a timer to supervise when resources in source eNodeB and resources in source Serving GW (if relocated) shall be released. If MME is relocated, the target MME also starts a timer to supervise resources for indirect forwarding at the target side. The length of the timer is chosen carefully to ensure that such resource releasing only happens after those resource is not needed. On the expiry of such timers, step 21 and 25 are performed, respectively, such resources at the source side and target side would be released.

The MME update bearer-related information, such as target eNodeB addresses and tunnel endpoint identities, is used for downlink traffic of those accepted bearers (by target eNodeB) to a Serving GW (step 17). For those non-accepted bearers, MME releases them by triggering the bearer release procedures. If the Serving GW is relocated, the target Serving GW update bearers information with PDN GW (Step 18). Such information includes target Serving GW assigned Serving GW addresses and tunnel endpoint indenters for each bearer (for both accepted and non-accepted). Then, PDN GW switches downlink data transfer to a new path, i.e., starts sending down link data to the target Serving GW, and sends “end marker” packets on old downlink path once the switch is completed, to notify source eNodeB that there is no more downlink packet coming.

The Serving GW sends it assigned Serving GW addresses and tunnel endpoint indenters for uplink traffic to the target eNodeB (step 19). If the Serving GW is not relocated, this Serving GW sends “end marker” packets on old downlink path once it starts sending downlink packets to the target eNodeB.

The terminal initiates tracking area update procedure at target side if it detects that it has entered a tracking area which is not in its stored TAI list, or other conditions for initiating this procedure are met (step 20).

The source eNodeB may cancel the handover procedure at any time before commanding terminal to move to target cell. And if none of the bearers requested by the source side could be established at the target side during the preparation phase, the target eNodeB rejects this handover.

(5) X2 handover with Serving GW relocation

During X2 handover, MME is unchanged but Serving GW is changed to a new one, please refer to Fig. 5.27. The handover preparation and execution of this procedure are same as those X2 handover without Serving GW relocation.

The MME decides that Serving GW should be relocated to a new one after receiving path switch request from the target eNodeB (step 1). Afterward, user plane tunnels for each bearer at the target side are established from step 2 to 5, which is similar to that of S1 handover with Serving GW relocated procedure, as well as the same downlink packets handling of the PDN-GW for both procedures.

Uplink data from terminal may be received and forwarded to source Serving GW, by target eNodeB, before receiving path switch acknowledgment (step 7). Once received, the target eNodeB sends uplink user data to target Serving GW. This is why the IP connection between target eNodeB and source Serving GW must be presented for this procedure. Otherwise, S1 handover is used instead.

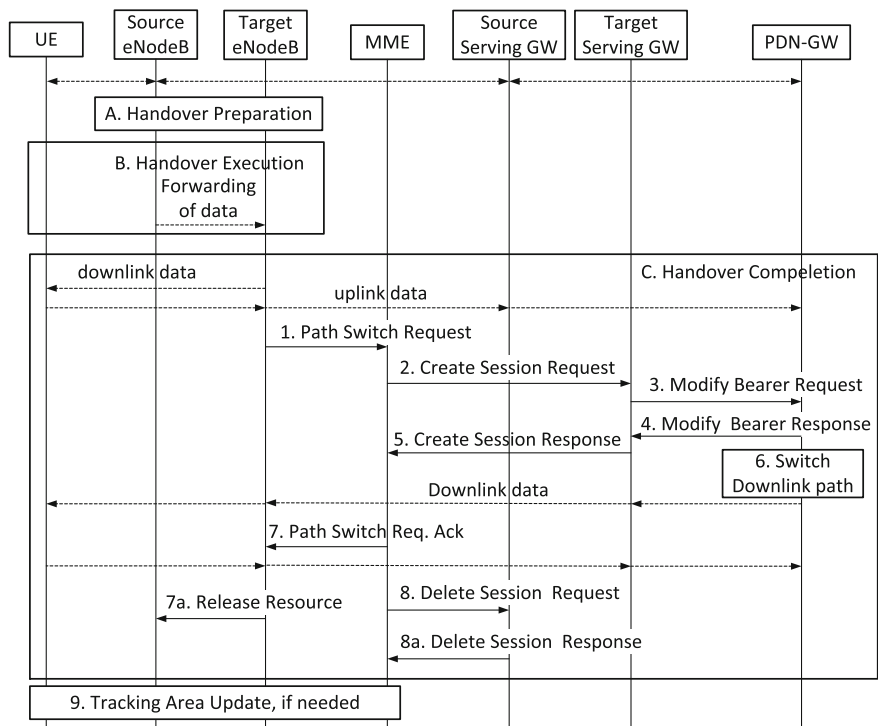


Fig. 5.27 X2 handover, with Serving GW change

(6) Inter-RAT Handover Between E-UTRAN and GERAN

The overall principle when designing the inter-RAT handover mechanism is that impacts/changes to previous generation radio networks, such as GERAN and UTRAN, must be minimized. The overall principle could be broken down to concreted principles [24], and those considered very important for understanding inter-RAT Handover procedure are listed as below:

- As an network-controlled handover, the inter-RAT HO is controlled by the source access system, e.g., the source access system decides when and to where to handover, decides the starting of the preparation phase, provides the necessary information to the target system in the format that the target system requires (i.e., the source system adapts to the target system), and decides when to executes the actual handover;
- Radio resources are prepared in the target 3GPP access system before the terminal is commanded by the source 3GPP access system to change to the target 3GPP access system. If RAN level interfaces are not available to enable this preparation, control interfaces in CN level should be used, e.g., an interface between 2G/3G SGSN and corresponding MME/Serving Gateway.
- The target access system should be responsible for giving exact guidance to the terminal on how to make the radio access there, such guidance this includes radio resource configuration, target cell system information, etc., during the handover preparation phase, and the transportation of such guidance to terminal should be completely transparently through the source access system.
- In order to avoid or mitigating the loss of user data, data forwarding mechanism can be used, until the GGSN/PDN GW determines that it can send downlink user data directly to the target system.
- The handover procedure should not require any UE to CN signaling in order for user data to be transported in the target system. That is, related information such as the security context, UE capability context, and QoS context is transferred (or translated if needed) within the network, between source and target system.
- Similar handover procedure should apply for handovers of both real-time and non-real-time services.

Generally speaking, most of those principles are also correct for intra-handover cases. Bearing those principles in mind, readers are encouraged to refer to detailed procedure documented in TS23.060 [16] and TS23.401 [15]. Since costs for upgrading or modify existing system are prohibited high; thus, the principle that the new system adapted to old system is taken when designing procedures and messages for inter-system mobility procedures. For example, the new system is responsible for using the existing message defined in old system to interact with old system, and to translate parameters that used in new system to that old system could understand before sending those parameters to old system.

5.5.3 Handover Control for CSFB and SRVCC

Voice service in LTE network could be provided via either circuit-switched fallback (CSFB) or single radio voice call continuity (SRVCC). For CSFB, the mobile terminal needs to fallback to CS domain for a mobile-terminated voice call arrived or the mobile terminal wants to originate a voice call. For SRVCC, the voice call is anchored in IMS domain; however, if a mobile terminal moves from LTE or HSPA coverage into GERAN/UTRAN coverage, the leg of this ongoing call of the terminal would move to CS domain, via a SRVCC procedure. However, the call leg in CS domain could be moved to IMS domain via a return SRVCC procedure, if possible. The handover control methods for above procedures are briefly introduced below.

(1) Handover control during CSFB procedure

Within CSFB deployed network, if a MT call arrives while the called mobile terminal is camped in LTE network (either in connected mode or idle mode), then a circuit-switched fallback procedure to 2G/3G CS domain is performed. An example procedure is shown in Fig. 5.28.

After inquires the HLR, the gateway MSC knows current registered MSC of called party and the send IAM messaged to this MSC. A SGs Paging message is then sent to MME by this MSC over SGs interface (step 1). If there exists signaling connections between MME and terminal (i.e., terminal is in connected mode, refer to Step B), then MME generates a CS service notification message and sends it to terminal via this connection. Otherwise, MME pages terminal, and terminal response with extended service request message, and MME step UE context in eNodeB (step 2 to 4), refer to Step A).

Since known that there is a MT CS call arrived for this terminal, MME sends UE context modification message with CSFB indicator to eNodeB (step 8). Based on local information, eNodeB knows whether both network and terminal support PS handover from LTE to 2G/3G network or not (at point a). If PS handover is supported, normal PS handover procedure (which contains preparation and execution phase) is initiated by eNodeB, by sending handover required message to MME (step 9). Otherwise, the eNodeB uses CCO procedure to move terminal to target 2G/3G cell (this is not shown in this figure).

Upon received handover command From E-UTRAN, mobile terminal connects to a cell in the target RAN. After that, i.e., at point b, if the 2G/3G RAN or terminal cannot support dual transfer mode, suspend procedure is initiated by terminal to suspend PS services. If LAI of current camped cell is different than that it stored, then mobile terminal send location area update request with MT CS flag to network, the MSC who receiving this request tries to continue the LAU procedure (step 11) to see whether this terminal allowed to access in this location area, if allowed, MSC establishes MT CS call (if this MSC is not the MSC who send the SGs Paging, then roaming retry procedure would happen before this establishment). If the location area update procedure is not performed, the mobile terminal sends paging response

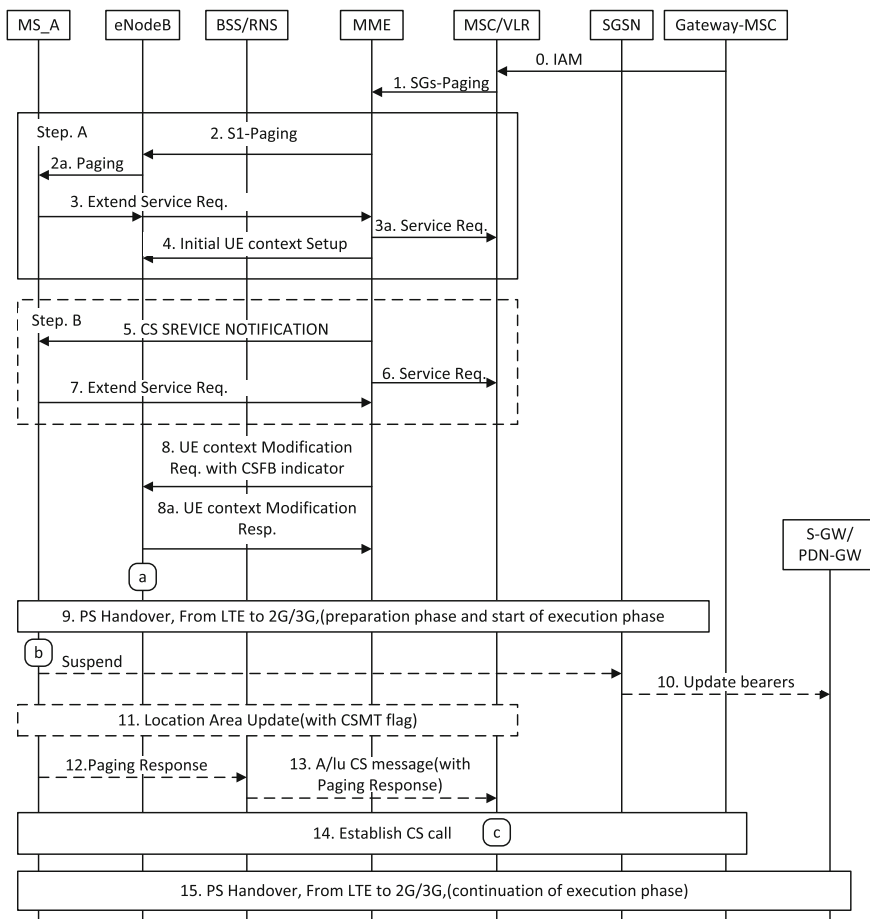


Fig. 5.28 An example CSFB procedure with PS handover supported

to network via GERAN/UTRAN (step 12), and then MSC establishes MT CS call after receiving the response (step 14). Then, mobile terminal performs remaining steps of PS handover execution phase (step 15), e.g., perform routing area update procedure.

At the decision point b, if the PS handover is not supported, the eNodeB instructs terminal to move/reselect to either a target GERAN cell or a target RAT (to which cell or to which RAT is specified by this eNodeB) via corresponding RRC message with CSFB indicator, then terminal either sends paging response to network or initiated location update procedure once it has camped on target cell of the target RAT.

Once the CS call is terminated, mobile terminal will either remain in GERARN/UTRAN and follow normal mobility management procedure or reselect to LTE, which the latter is a recommend terminal behavior.

(2) Handover control during SRVCC procedure

Within SRVCC deployed network, voice calls are anchored in IMS layer, i.e., mobile terminals use PS domain rather the CS domain to originating/receiving voice calls. However, if mobile terminal moves into radio coverage that unable to support IMS call, the ongoing IMS anchored call needs to be continued in CS domain via SRVCC procedure. An example procedure is shown in Fig. 5.29.

After receiving measurement report, the eNodeB decides (step 2) that a PS handover is needed (and it is supported by 2G/3G radio network and terminal) and figures out that the imminent handover is SRVCC handover, i.e., this is a PS + CS handover, then the eNodeB notify this handover type to MME, as well as target side information such as target cell identity (step 3).

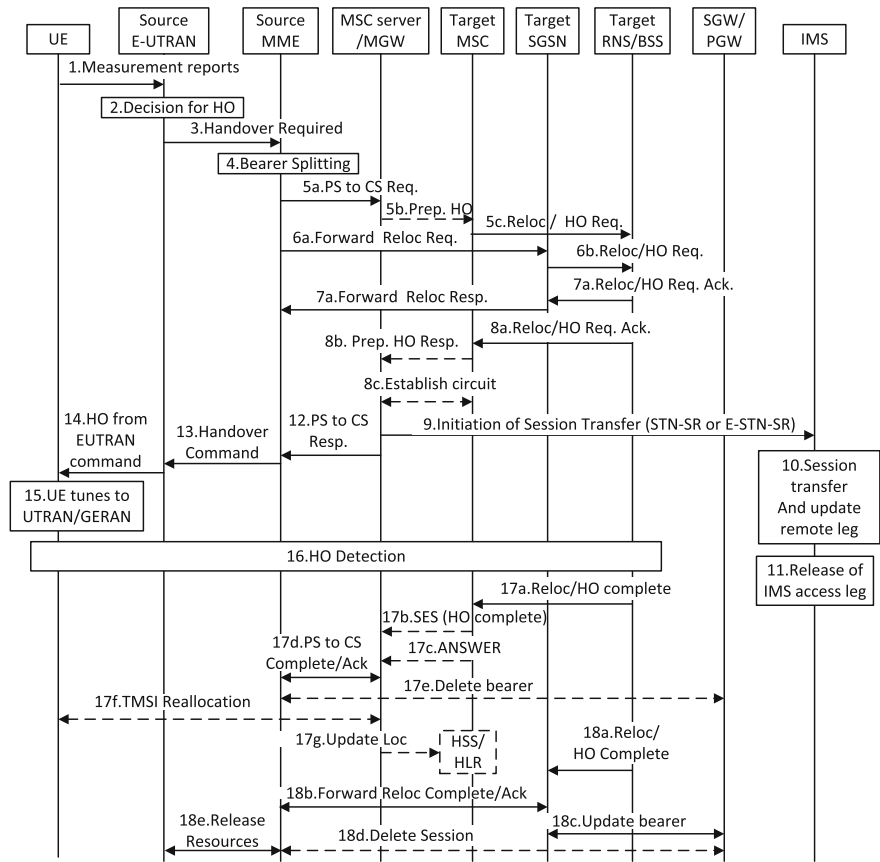


Fig. 5.29 An example SRVCC procedure, with PS handover supported

The MME figures out which data bearers are CS related and which are PS related (i.e., bearer splitting) by reading bearer's QCI value from locally stored UE Context, e.g., QCI = 1 means this bearer is a bearer for CS voice (step 4).

Then MME notifies MSC Server and SGSN in parallel (step 5a, 6a, respectively) that handover is required with target side information, CS service-related bearer information and PS service-related bearer information, respectively. The MSC Server interacts with target MSC to prepare CS circuits and radio resources in target cell and also initiates the Session Transfer procedure to IMS domain entity, to transfer this IMS Call leg to CS leg (step 10, 11).

After both MSC server and SGSN complete their handover preparation, the MME issues a handover command to source E-UTRAN (step 13). The mobile terminal tuned to target GERAN/UTRAN and sends handover complete to network. Then, PS handover and CS handover complete after necessary interactions to release resources within source side, and possible location update procedures (step 17). Then, the target side confirms this handover and updates bearers accordingly (step 18a, 18b and 18c). The source side then releases resources after got the confirmation (step 18d, 18e).

(3) Handover control during return SRVCC procedure

For both voice call that anchored in IMS layer continues in CS domain after SRVCC procedure and voice call originated from GERAN/UTRAN in CS domain, those calls could be moved and continues in IMS domain if terminal and network support. An example procedure for this return SRVCC (also named as CS to PS SRVCC) procedure is shown in Fig. 5.30. The source side is GERAN without DTM and PS handover support, and target side is E-UTRAN or UTRAN with HSPA (the data rate of HSPA meets IMS voice requirement).

After receiving handover required message including Target Id, CS to PS indication, from the RNC/BSC, the MSC Server retrieves PS registration information (e.g., terminal's latest temporary identity) from terminal, selects target MME/SGSN based on Target ID, and then sends CS to PS handover request to this selected MME/SGSN (step 1, 1a, 3), and session transfer notification to IMS domain (step 2).

If no UE context stored, the target MME/SGSN would fetch UE context from old MME/SGSN by using the latest temporary identity, before requesting target radio network to prepare/allocate resources for this handover (step 4, 5, 6). Once preparation successes, target MME/SGSN response back to MSC Server (step 7). The MSC Server commands terminal to perform CS to PS handover (step 8) and provides the IP address/ports and selected codec for the IMS media anchoring to this terminal at the same time. The MSC Server also requests IMS to switch the media path switch to the IP address/port of the terminal on the target access (step 9).

After received handover confirmation from terminal, eNodeB/NodeB notifies target MME/SGSN, then MME/SGSN confirm this handover competition to MSC server and update PS bearers with SGW/PDN-GW (step 10, 11, 11a, 12).

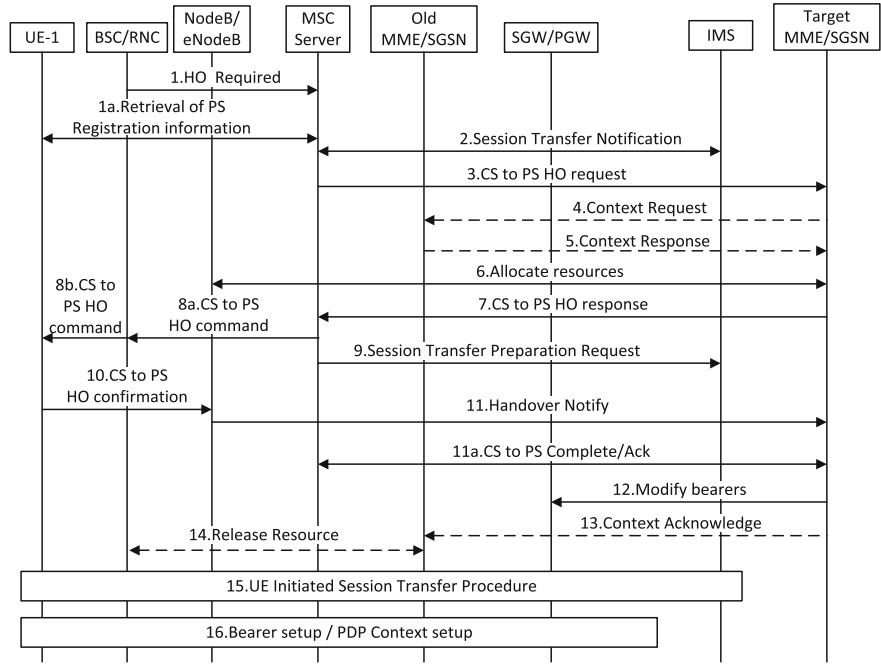


Fig. 5.30 An example of return SRVCC procedure. Reprinted from 3GPP TS23.216 [50]

The terminal then initiates session transfer procedures, and dedicated bearer or PDP Context for voice would be established to carry voice media with QoS guarantee (step 15, 16). The voice media data then would be carried by dedicated bearer rather than carried by default bearer temporarily. The terminal initiates TAU/RAU after this handover, and the target MME/SGSN knows that it is related to CS to PS handover.

5.6 Security Mechanism

5.6.1 Security Mechanisms in GSM, UMTS, and LTE

(1) General Idea of GSM security

Within the scope of cellular network, mobility management interacts with security mechanism at various aspects. This section will give a brief introduction to the security mechanisms designed for the cellular network.

Some 1G mobile networks have little security mechanism, such as NMT, while most 1G mobile networks have no security mechanism at all, e.g., analog-AMPS.

The following security mechanisms have been introduced into GSM system: (a) user authentication and service authorization, (b) confidential protection for

private user information (e.g., user's identity, user data, location, and movement track) against exposure to unauthorized network entities, third party, and procedures.

The major breakthrough of GSM on security aspects is that using a smart card, which is a physical entity inserted into the MS, to store user's identity, shared secret key, and algorithms used for security purposes, and other subscription data. In GSM system, this smart card is called SIM card, which consists of storage and a SIM Application used for GSM security-related procedures. This mechanism is kept and enhanced in UMTS and LTE system [43], while in those systems it is named as UICC card, with corresponding UMTS and LTE security parameters and algorithms enhanced, and the USIM Application residing on the UICC is responsible for UMTS/LTE security-related procedures [43]. For backward compatibility reasons, UICC may have SIM-ME interface for MSs only supporting GSM authentication. The key stored in SIM card is shared with the authentication center (AuC) in the user's home environment, on the per subscriber basis, and the key will never leave SIM card and AuC for security reason.

The GSM authentication procedure is based on challenge/response mode. The MS provides its IMSI to the AuC firstly, then AuC generates a network challenge RAND, retrieves the key K_i by the IMSI, and uses the RAND and K_i as input to A3 and A8 algorithm to derive an expected user response SRES, and a ciphering key K_c , respectively. Then, the triplet (RAND, SRES, K_c) are sent to the serving MSC/VLR of the MS and stored in MSC/VLR. The RAND is then sent to MS, the SIM card will use its stored K_i and this RAND as input to A3 and A8 algorithms to derive the SRES and K_c , respectively. After that, the MS sends the SRES to the serving MSC/VLR, the MSC/VLR will compare received SRES from MS with SRES from AuC, if identical, then this MS is authenticated by the network. See Fig. 5.31a. The K_c will be passed to the base station by serving MSC/VLR. After a command issued by the MSC/VLR, the user data ciphering begins between MS and base station, by using the A5 algorithm and this K_c . A3/A5/A8 algorithms mentioned above are not any kind of specific algorithms, actually they are referred as requirements for algorithms being used in above scenarios. Typical examples of A3/A8 and A5 algorithms are COMP128 and KAUSMI. The access authentication procedure in GPRS system is the same; with the exception that the authentication is performed by SGSN.

In order to protect user's identity, user's location privacy and prevent user being tracked by malicious people, a temporary identifier is used to identify a user on the air interface. The temporary identifier (i.e., TMSI) is allocated to the MS when necessary and the correlation with user's IMSI is maintained by the MSC/VLR. Normally, the MSC/VLR can allocate a new TMSI to the MS when it wishes to do so. The TMSI contains information to find the MSC/VLR who allocates this TMSI. When the MS identify itself with this TMSI, the base station can find correct MSC/VLR for this MS. When MS moves to another MSC/VLR, it will get a new TMSI from the new MSC/VLR. The old TMSI could be allocated to another MS. Since the correlation between TMSI and IMSI is changed dynamically, it is very hard to track the user. This idea of using temporary identity at air interface is also kept in GPRS/UMTS and LTE system.

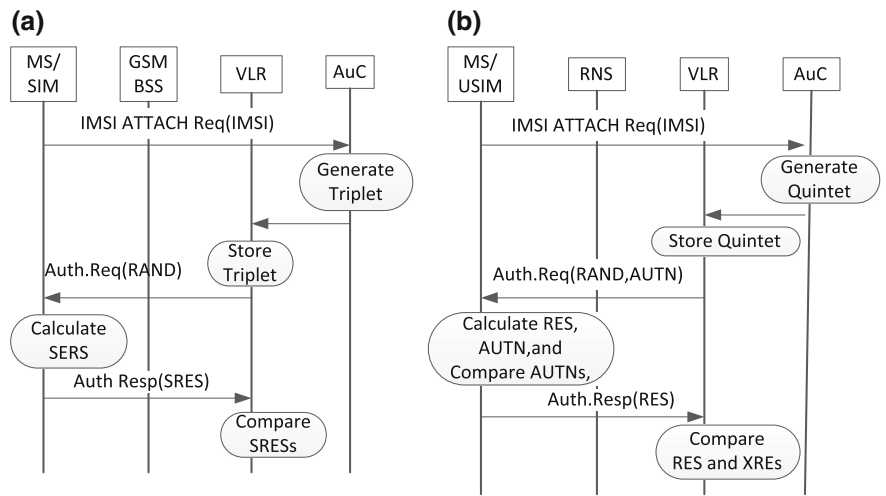


Fig. 5.31 GSM and UMTS authentication procedures. **a** GSM authentication procedures **b** UMTS AKA procedures

However, the GSM system is still confront with following security threats, such as SIM card attacking, user identity masquerading attacking and signaling network attacking. Following aspects are major flaws led to possible successful attacks mentioned above, respectively, one-way authentication (i.e., the SIM card could not authenticate who send the challenge, the MS could not authentication the network), no security mechanisms for links between network nodes (e.g., base station and base station controller, that means Kc and SRES are transported within the network in plain text). Even more, the GSM system has no integrity protection for both user and signaling data.

(2) Security Architecture of UMTS and LTE

Since Rel-99, the security architecture for 3G has been introduced. Within this architecture, the following five security feature groups are defined: (I) network access security; (II) network domain security; (III) user domain security; (IV) application domain security; and (V) visibility and configurability of security. This security architecture with five security feature groups are reused in LTE, with following three major enhancement: network access security extended to protect links between mobile equipment (ME) and serving network, network domain security extended to protect links between access network and SN, and the home environment (HE) also needs certificate the SN [44].

Please find the overview of the LTE security architecture in Fig. 5.32. Detailed description of those five features could be found in [45]. Following of this book will focus on security feature group (I), who has tight relationship with mobility management.

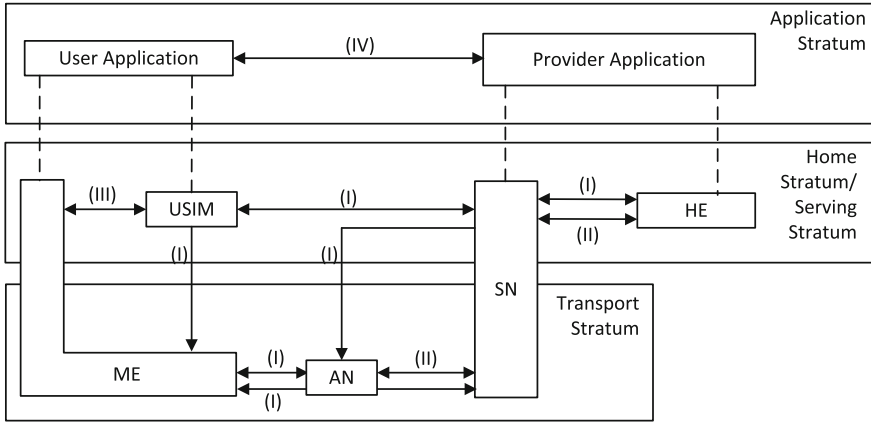


Fig. 5.32 LTE security architecture. Reprinted from 3GPP TS33.401 [45]

(3) Authentication and Key Agreement, Ciphering, and Integrity Protection in UTM/LTE

Within the network access security feature group of the UMTS, the most notable improvement is authentication procedure, which is named authentication and key agreement (i.e., UMTS-AKA). In order to achieve maximum compatibility with the current GSM security architecture and facilitate migration from GSM to UMTS, the UMTS-AKA is composed of a challenge/response protocol identical to the GSM subscriber authentication, and a key establishment protocol combined with a sequence number-based one-pass protocol for mutual authentication between MS and network [45]. Compared with GSM, the authentication vector in UMTS is named Quintet, it contains (a) a network challenge RAND, (b) an expected user response XRES, (c) a cipher key CK, (d) an integrity key IK, and (e) a network authentication token AUTN. For the authentication procedure, not only the MSC/VLR and SGSN need authenticate the MS, but also the MS authenticates the network by compare received AUTN with locally calculated AUTN. See Fig. 5.31b. This allows the mobile station to detect a false base station. Detailed procedures could be found in [45]. In UMTS, the access link signaling integrity protection is introduced, which allows the NAS signaling to be protected between MS and the radio network controller as well. The user plane data ciphering and integrity protection are performed between MSs and RNC.

The idea of UMTS-AKA is reused in LTE, with name changed to EPS-AKA, and the EPS-Authentication Vector reduced to four elements (i.e., K_{asme}, RAND, AUTN, and XRES). Moreover, the network access security in LTE is enhanced by introducing a sophisticated key hierarchy, key distribution, and key derivation scheme [45, 46]. With this enhancement, the key separation for serving network is achieved by using serving network identity as additional input when generating keys. Further, the AuC in user's HE does not generate keys used directly in encryption and integrity protection. The AuC only generates a K_{asme}. The K_{asme}

is then distributed to the access security management entity (i.e., MME) to generate keys used for NAS signaling encryption and integrity protection, and a Kenb when necessary. This Kenb is distributed to eNodeB in order to generate keys used within access network for encryption and integrity protection for both RRC signaling and user plane data. Thus, the key separation between networks and key usage are fulfilled.

For ciphering/encryption and integrity protection discussed above to take into effective, explicit signaling procedures to command the MS to start ciphering/encryption and integrity protection are needed. For UMTS/LTE system, ciphering/encryption and integrity protection are always started with selected algorithm, which is indicated to MS by network during the signaling procedure. There is an algorithm referred as “NULL,” if selected, there is no protection. Readers who interested in security algorithms may refer to [45].

(4) A brief comparisons among 2G/3G/4G Securities

Comparisons among security aspects of GSM, UMTS, and LTE could be found in Table 5.8.

Table 5.8 Comparisons among security aspects of GSM/GPRS, UMTS, and LTE

	GSM/GPRS	UMTS	LTE
Authentication method (and its name)	Network authenticate MS (GSM authentication)	Mutual authentication between UE and network (UMTS-AKA)	Mutual authentication between UE and network (EPS-AKA)
Authentication vector	Triplet (RAND, SRES, Kc)	Quintet (RAND, XRES, CK, IK, AUTN)	EPS-authentication vector (Kasme, RAND, AUTN, XRES)
Key derivation function/algorithm	A3 and A8 used derive SRES and Kc, respectively	Milenage, used to derive the Quintet	Milenage, used to derive the EPS-authentication vector
Key length (cipher and integrity protection)	64bit 128 bit (from R9 + GSM onward)	128bit	128 bit key is used now, algorithms already support 256 bit
Ciphering algorithm	A5 for CS data, GPRS A5 or GEA for packet data	NULL, KASUMI, SNOW 3G	NULL, NOW 3G, AES, ZUC
Objects of ciphering	User data transmitted between mobile station and base station	User data and Signaling transmitted between mobile station and RNC	User Plane between UE and eNodeB, RRC and NAS signaling, are encryption protected
Integrity protection algorithm	–	NULL, KASUMI, SNOW 3G	NULL, SNOW 3G, AES, ZUC
Objects of integrity protection	–	Radio access signaling link (the NAS signaling is protect between mobile station and RNC)	RRC and NAS signaling

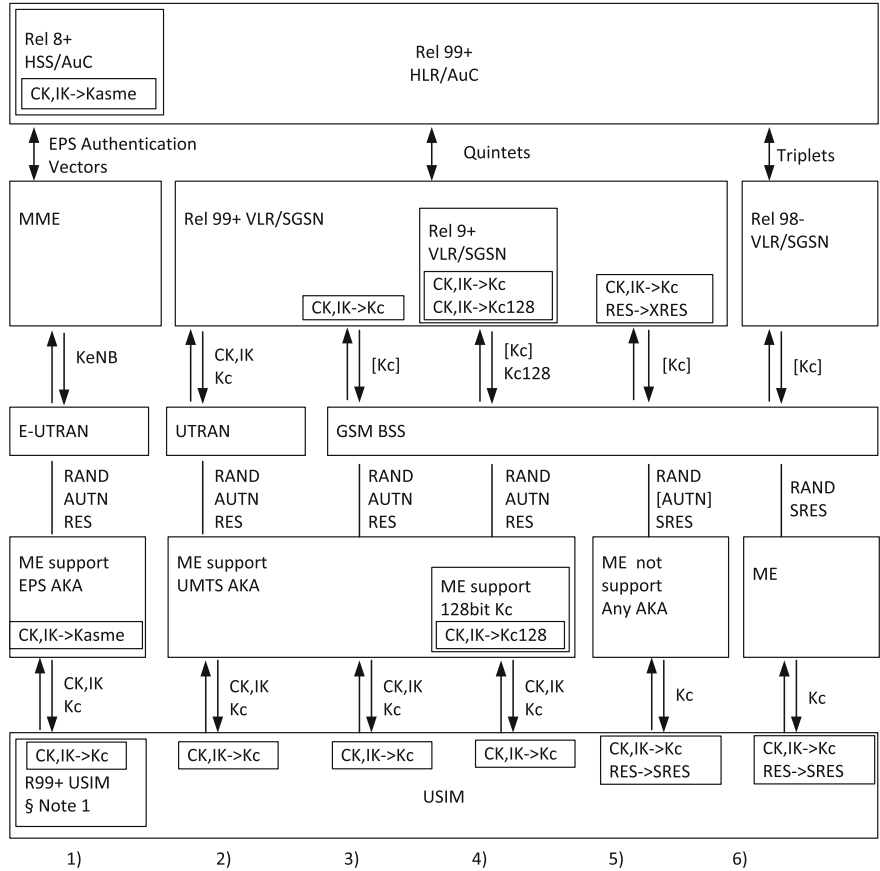


Fig. 5.33 Scenarios for UMTS/LTE subscribers in a mixed network

(5) Scenarios for UMTS/LTE subscribers in a mixed network

Depending on network capability (e.g., different VLR/SGSN, HLR/AuC, access network are deployed according to different releases), mobile equipment capability, and SIM or USIM card inserted; there are some different combinations in real world. Figure 5.33 gives six different scenarios (i.e., six columns with different combinations) assuming USIM is used.

For (1), please look from bottom to top at the most left column, the Rel-99 + USIM is inserted into a LTE UE, and the EPS-AKA must be performed when UE accessing via E-UTRAN (with MME and AuC involved). The AuC and the LTE user equipment should support converting the (IK, CK) to the Kasme.

For (2), and (3), (4), the MS is capable of UMTS-AKA and accessing 3G Core network (i.e., the VLR/SGSN is Rel99 or onwards) via UTRAN and GSM BSS,

respectively. For (2), the UMS AKA must be performed when UE accessing via UTRAN. During the UMTS-AKA procedure, AuC sends quintets to R99+ VLR in those 3 scenarios. The R99+ VLR converts the (IK, CK) to the Kc and sends it to GSM BSS for ciphering and integrity protection for (3) and (4). The difference is that the VLR/SGSN is Rel9 or onwards in (4), so 128bit Kc is used in this case.

For (5), the USIM use SIM-ME interface to interact with the MS since the MS supports neither UMTS-AKA nor EPS-AKA. In this case, the MS is accessing 3G Core network via GSM BSS. So, the UMTS-AKA is performed with following modifications, both VLR and USIM should convert quintets to triplets and use triplets.

For (6), the Rel98-(or earlier)VLR/SGSN is deployed, which means the core network is 2G, so regardless MS's capability, the GSM authentication is performed, where USIM should convert quintets to triplets.

For GSM subscriber using SIM card, scenarios for mixed network could be found in [44].

One important decision made by 3GPP is that in order to achieve a reasonable security level for LTE network, the LTE network will not grant access to MSs with SIM card, thus known GSM and SIM card-related security threatens problems are kept away from LTE network.

5.6.2 Security Mechanisms and Mobility Management Interactions

The operator networks are deployed and evolved gradually, e.g., newly added network entities may have new features and interact with existing network nodes smoothly. Network entities with different capabilities (e.g., comply to different standards releases) must interoperate in a compatible way, to support the user mobility and service continuity within the mixed network environments. As introduced above, 2G/3G/4G systems have their own security mechanism to achieve certain security level, and each system should maintain its security level. Thus, intersystem mobility should avoid spreading vulnerability of a source system into a target system not susceptible to the same vulnerability; otherwise, higher security level of target system would be downgraded. All those should be considered for interactions of security mechanisms and mobility management. Two tightly related aspects are discussed as below.

(1) Security-related Signaling Optimization for Mobility

Performing a new authentication procedure each time a MS enters the network is time-consuming. For a MS initially attaches to the network, it should be ok to do so. However, it might be not desirable for MSs that handover within the same access, or handover from one radio access to another (e.g., from LTE to UTRA). In order to reduce delay caused by full authentication procedure, following methods are taken. The AuC will send multiple authentication vectors to a mobility management node

(e.g., VLR, SGSN, and MME) during one signaling interrogation procedures. The mobility management node will store those authentication vectors for future use and transfer authentication vectors to another mobility management node if requested during inter mobility management nodes mobility. Such optimization also saves signaling between mobility management nodes and AuC.

For example, the VLR normally gets five triplets from AuC when one MS attaches to this VLR with IMSI. The VLR picks up one triplet to authenticate this MS, if succeed, this triplet is marked as used. All those triplets are stored in VLR for hours or longer even after this MS has been detached. Assuming this MS (re) attaches with the TMSI allocated during previous attach procedure to the same VLR, the VLR then uses this TMSI retrieves locally stored triplets and picks up an unused one for authentication. Further, if the MS moves to another VLR, this new VLR will first request triplets from previous VLR by using received TMSI. If available, the new VLR picks up an unused triplet to authenticate this MS, if necessary.

For intersystem mobility between 4G and CDMA2000, mobile station performs pre-authentication (or pre-registration) with target system via source access and a signaling tunnel which between the source system and target system (i.e., S101 interface). For example, a pre-authentication done is prior to MS actual handover from 4G to CDMA2000 access. With pre-authentication procedures, handover delay for 4G to CDMA2000 mobility would not contain time spent on security procedures.

(2) Security Context handling for Inter-3GPP Access Mobility Scenarios

Security context contains security-related parameters, such as authentication vectors and keys for ciphering. In order to obtain service from 2G/3G/4G systems, security context must be established within MS and network entities (e.g., MME, VLR, SGSN). The full authentication or AKA procedure is normally used when the MS initial attaches to network, if there are no valid stored security contexts that established during previous registration procedure. Security context can also be established by mapping a security context from source system to a security context used in target system during mobility procedures, e.g., a mapped EPS security context is established in MME by mapping UMTS security context when MS handover from 3G access to LTE access. Obviously, the handover delay is reduced, since security context is established by mapping rather than by performing a full authentication or AKA procedure.

We take mobility between 3G and 4G as examples to illustrate how security and mobility interaction works, please refer to Fig. 5.34.

In step (1), the MS without any store security context initially attaches to LTE network with performing a full EPS-AKA procedure, MS and MME will establish a *native* ESP security context, respectively. That is, the MS and MME would have the same *native* K_{asme}.

For LTE to 3G mobility shown in Fig. 5.34a, it is MME's and MS's responsibility to map K_{asme} to CK's and IK's (step 2)). The mapping is based on a

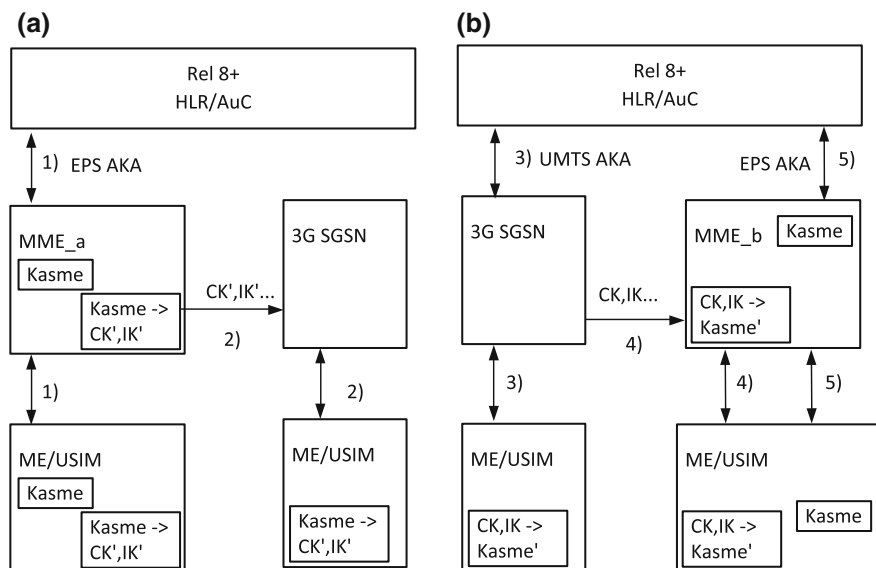


Fig. 5.34 Two scenarios for mobility and security interaction. **a** LTE to UMTS mobility. **b** UMTS to LTE mobility

cryptographic key derivation function having the property that it protects the source context from mapped target context [18]. This assures that if attackers compromise a mapped context they get no information about context from which it was mapped.

In order to limit the distributed scope of EPS security context within EPS network and support interworking with existing 3G network (i.e., without upgrading SGSN), the principle that “LTE side adapted to UTRAN/GERAN side” for 3G/2G to 4G mobility procedure is adopted. That is, for 3G to LTE mobility shown in Fig. 5.34b, if there is no available EPS security context, which is obtained and cached/stored by MS and MME in pervious network access procedure, it is MME’s and MS’s responsibility to do security context mapping (step 4), where MS moves to LTE sometimes after a UMTS-AKA is succeeded in step (3). The MME receives CK and IK from SGSN and maps them to Kasme. The MS also needs map CK and IK to Kasme. However, such mapping of security context can only achieve protection in one direction, if source context has already been compromised, then the mapped context will inherit this property [18].

In order to protect target system against threatens from possible compromised source system, especially the target system has higher security level, for example, it is highly recommended that to activated cached *native* EPS security context (if available) after mobility from 2G/3G, where *mapped* security context is used during the mobility procedure. If there is no *native* EPS security context in this case, a full EPS-AKA procedure should be initiated by network as soon as possible once the mobility procedure finished. This is shown as step (5) in Fig. 5.34b.

(3) Summary

With above mechanisms, the delays of mobility management procedures such as attach, location update, intrasystem handover, and intersystem mobility are reduced greatly and service continuity is achieved as well, while without compromising the security of the system at the same time.

Security mechanisms and mobility management interaction is a very complicate topic due to user mobility, mix network environment, and diversified MS capabilities. It is not possible for this book to cover every interaction scenarios and go through every detailed aspects of this topic. Readers who interested in this topic are recommended to read TS33.102 [44], TS33.401 [45], TS24.008 [29], TS 24301 [28] for details.

5.7 Interworking and Convergence Between WLAN and Cellular Networks

Broadband wireless access technologies, such as Wi-Fi, focused on the last mile of the fixed (and nomadic) wireless access within a local area, while cellular networks can offer mobile wireless access across wide areas at relatively low data rates and high costs. Thanks to its high access data rate, low cost and ease of use characteristics, the wireless local area network (WLAN, as defined by IEEE 802.11 standards) is becoming increasingly popular. Initially, the WLAN and cellular networks were considered to be two rivals, but soon, it turns out they complement each other in following five major characteristics shown in Fig. 5.35.

In order to benefits from both systems, many research works on interworking between broadband wireless access systems and cellular systems has done since 1998. The challenges faced by such interworking were raised from following aspects: interworking architecture, authentication, authorization, and accounting;

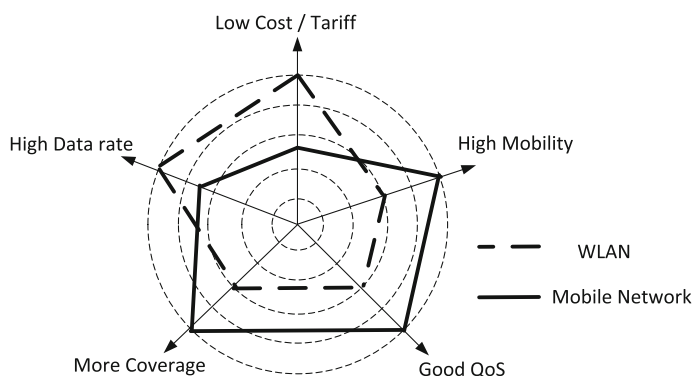


Fig. 5.35 Performance comparison of WLANs and mobile networks

security; location management; service continuity; handover control; roaming; and etc.

5.7.1 The Loose Coupling and Tight Coupling Interworking Architectures

Loose coupling and tight coupling are two interworking architectures that were proposed by the ETSI in 1998, where the interworking takes place between the high-performance radio LAN (HiperLAN) and the UMTS network (i.e., the 3G network) [47].

Within the loose coupling architecture (as shown in Fig. 5.36a), the user data are routed directly to and from the Internet by HiperLAN. The AAA-related procedure involves the Home Location Register (HLR), AAA servers located in 3G and HiperLAN network, and AAA Interworking Units (IWU). The AAA procedure is controlled by the 3G network operator and common billing for users that access the HiperLAN and the 2G/3G networks is achieved. The identifiers from the Universal Subscriber Identity Module (USIM) (if the terminal has one) or the Network Access Identifier (NAI) are used in corresponding AAA procedures, by using a USIM-centric IWU or a NAI-centric IWU for interworking, respectively. Within HiperLAN, mechanisms for security, mobility, and QoS can be deployed independently. No standardized method has been defined for mobility between the HiperLAN and the 3G network.

In the tight coupling architecture, the 3G core network controls the HiperLAN in a similar way that controls the 3G radio access network. A newly introduced IWU mimics the RNC, and the HiperLAN access point is considered a base station. The interfaces between the access point and the IWU, between the IWUs, and between the IWU and the SGSN are introduced as counterparts to the Iub, Iur, and Iu-PS interfaces, respectively, as well as the interface between the IWU and the RNC (as shown in Fig. 5.36b). Within this architecture, the mobility management mechanism that HiperLAN related is achieved at the data link layer, which is the same as that of the 3G networks. Mobility between HiperLAN access and 3G access is also supported. To fulfill this architecture, new interfaces, SGSN, and GGSN function modifications, and a new RNC-like IWU must be developed, which would require lots of standardization and development work.

5.7.2 I-WLAN Interworking Architecture

(1) UMTS and WLAN interworking defined by 3GPP

The 3GPP started the work on WLAN-3G interworking in its Rel-6 work and finished in Rel-8, and the output of HiperLAN-3G interworking was considered

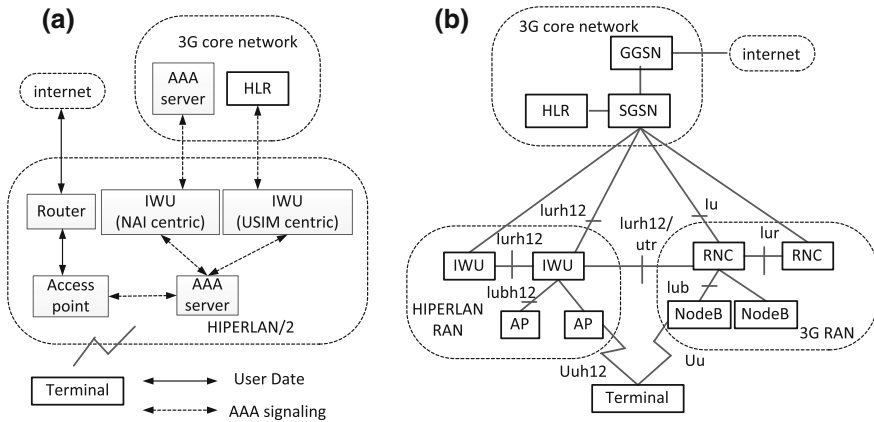


Fig. 5.36 Interwork architectures of HiperLAN/2 and 3G [UMTS] networks. **a** Loose coupling **b** Tight coupling

during this work (referred as I-WLAN). The purpose of the I-WLAN work is to extend 3GPP services and functionality to the WLAN access, as well as to other similar IP-based access networks. Six interworking scenarios are defined in an evolutionary path from entry-level service to full availability of 3G services, in TR22.934 [48]. Each of these scenarios provides an increasing level of service that is built upon the lower level scenarios. In the first scenario, common billing and common customer care are offered without any extra work based on the current specifications; 3GPP system-based access control and charging is added in the second scenario. In the third, users can access 3GPP PS-based services from the WLAN. In the fourth and fifth scenarios, service continuity and seamless service continuity are supported, respectively, with the difference that in the fifth scenarios the service interruption is not noticeable to the user during the handover from 3GPP to WLAN, or vice versa. The sixth scenario is access to 3GPP CS-based services via WLAN access; however, there is no actual work done since no desire to deploy this scenario. The 3GPP2 also defined five similar scenarios for interworking of CDMA and WLAN networks.

Two new procedures (and their corresponding architectures, see Fig. 5.37), WLAN Direct Access and WLAN 3GPP IP Access, are defined for I-WLAN. The former corresponds to the second scenario; it allows the terminal to access the Internet via the WLAN only, however, the AAA-related procedure is controlled by the 3G network. The latter corresponds to the third scenario, which allows the terminal to access the 3GPP PS services via the PS domain through the WLAN Access Gateway (WAG) and the Packet Data Gateway (PDG).

(2) Network selection

Multiple WLAN ANs may have interworking relationships with multiple PLMNs. To authenticate itself with its home public land mobile network (HPLMN), a WLAN UE (hereafter, we use “terminal” for simplicity) must select a

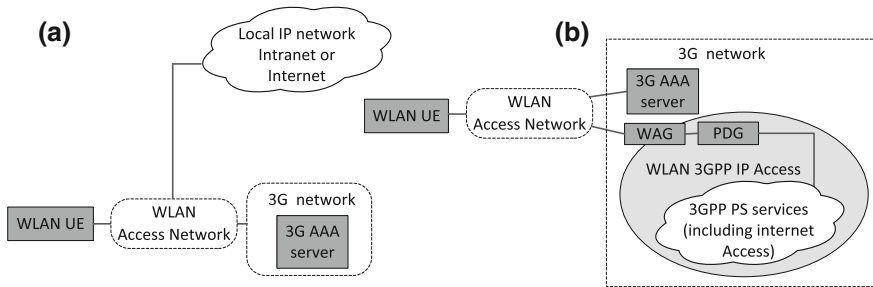


Fig. 5.37 Interworking architectures of I-WLAN. Reprinted from 3GPP TR22.934 [48]. **a** WLAN DIRECT IP Access **b** WLAN 3GPP IP Access

suitable WLAN AN (either in manual mode or automatic mode), to let AAA signaling reach its HPLMN via the selected WLAN [49].

In the manual mode, available Service Set IDs (SSIDs) and the IDs of their interworked PLMNs are presented by the terminal to the user for selecting one PLMN, once selected, then the terminal is associated with one WLAN AN that interwork with the selected PLMN, to proceed authentication procedure. In the automatic mode, the terminal is associated with the available WLAN AN whose SSID has highest priority in the I-WLAN identity preference list firstly and tries to perform authentication with the HPLMN via this WLAN AN. If the HPLMN cannot be reached via this WLAN AN, another WLAN AN selection procedure should be performed. The terminal determines the best match PLMN based on the network advertisement information received during the previous authentication procedure, combined with the user-controlled PLMN selection list for I-WLAN and the operator-controlled PLMN selection list for I-WLAN. Once PLMN is selected, the terminal is then associated with the WLAN AN whose SSID has the highest priority in the I-WLAN identity preference lists that interworks with selected PLMN, and authentication is performed between the terminal and the selected PLMN. Please note such lists are provisioned into USIM by management procedures.

(3) Security aspects

Security needs more consideration for I-WLAN systems because the WLAN itself cannot provide security protection for user data and privacy at the same security level that 3G achieved.

For the I-WLAN architecture, a mutual authentication procedure is performed between the terminal and 3GPP AAA server. The authentication methods used are the Extensible Authentication Protocol for GSM Subscriber Identity Module (EAP-SIM) and the Extensible Authentication Protocol for UMTS Authentication and Key Agreement (EAP-AKA) for terminals equipped with SIM or Universal Integrated Circuit Card (UICC) authentication termination points, respectively.

WLAN access authorization is used to verify whether WLAN access is allowed to a subscriber and to decide access rules/policies applied to this subscriber. This

procedure is performed when the authentication is successful, or can be combined with authentication procedure. For WLAN IP Access, the service authorization follows the above procedures, which is used to decide which types of services are allowed and not allowed by this user's home network, which is based on, for example, the W-APN (wireless access point name), the user subscription information, the visited network capabilities, and the roaming agreement.

The W-APN is the APN that is used in the I-WLAN architecture, and it is also used to select a suitable PDG, along with its authorization purposes. To protect the user data traffic and the 3GPP network, a security tunnel based on the Internet Key Exchange version 2 (IKEv2) protocol must be established between the terminal and PDG.

(4) Basic IP connectivity support for I-WLAN architectures

For WLAN direct IP Access, the terminal obtains a local IP address and the data traffic goes to the Internet via the WLAN only. In the WLAN IP Access case, the IP connectivity between the terminal and the PDG must be established before the terminal can access the 3GPP network-provided service via the PDG. The IP tunnels are established when the terminal is authenticated and authorized. The terminal obtains a local IP address from the WLAN access network, and the PDG allocates a remote IP address to the terminal. The PDG binds the remote IP address and the local address during the tunnel establishment procedure. The WAG enforces the forwarding of the tunneled data packets that are sent from the authenticated terminal to the specified PDGs. The PDG encapsulates and de-encapsulates the user service data packets, and routes the different users' service data to and from their corresponding data networks. The WAG and the PDG also perform charging information collection and reporting, QoS and operator policy enforcement, and data packet filtering functions.

(5) Service Continuity Support for the I-WLAN Architecture

Dual Stack Mobile IP (DSMIP), the host-based mobility (HBM) management protocol, was chosen by the 3GPP as a basis for achieving service continuity and seamless service continuity, i.e., to maintain the ongoing 3GPP PS-based services with minimal impact on the user's perceived QoS when the terminal changes access networks (e.g., between WLAN and 2G/3G access).

The most important function used to achieve service continuity is called *IP Address Preservation*, which ensures that the terminal's home address remains unchanged when the terminal changes access networks from 2G/3G to WLAN, or vice versa. An enhanced architecture is shown in Fig. 5.38 that supports HA discovery, security DSMIP signaling exchanges, and handover between WLAN and 2G/3G access while maintaining the service continuity.

For secure DSMIPv6 signaling exchanges, such as binding updates signaling, the terminal establishes a security association with the home agent. The HA's IP address is then made known to the terminal by one of the following methods; a static configuration in the terminal, a DNS query, delivery to the terminal in the

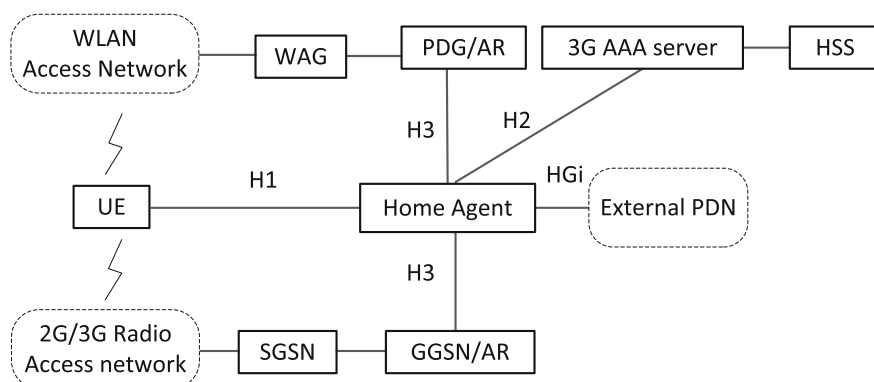


Fig. 5.38 Interworking architectures of I-WLAN with mobility support. Reprinted from 3GPP TS23.327 [51]

Protocol Configuration Option (PCO, see TS23.327 [51]), or delivery by the PDG within the IKEv2 configuration payload. Which one is used depends on the deployment of underlying IP network.

Before performing DSMIPv6 registration over any PDN connection the terminal must determine whether it is on its home link, by comparing its IPv6 prefix associated with a specific access interface and the home network prefix (HNP) allocated to the PDN connection [21]. If the two prefixes match, then the terminal determines that it is the home link for this specific PDN over the access interface. The terminal gets the HNP associated with a specific PDN from either IPsec security association bootstrap or PCO received in the 3GPP attach procedure. The terminal gets the IPv6 prefix associated with a specific access system interface via the IP address allocation mechanisms applied in that access system.

As a preferred implementation option, the HA functionality can be co-located with the GGSN. In this case, the GPRS system is configured as the home link, and the terminal gets the IP address prefix from the GPRS system and uses it as a home address. The HA can also be co-located with the PDG, and the 2G/3G access system may still act as the home link. In such case, the terminal gets the IP address prefix and uses it in the same way that it does in the previous case. In both cases, DSMIPv6 user plane encapsulation over the GPRS system is not used; the user plane data are still carried by the GTP tunnel, to minimize packets encapsulation effects of 3GPP system.

When the UE moves between 2G/3G access and WLAN access, its HA remains unchanged, and the GGSN can ensure that the allocated home IP address and prefixes within the new access are the same as those allocated to this UE within the previous access. Thus, the continuity of UE's IP connectivity in the interaccess system change case is achieved.

5.7.3 EPC-Centric Interworking Architecture

From their Rel-8 work, the 3GPP began research and standardization work on providing IP connectivity using non-3GPP access to the evolved 3GPP packet-switched domain, i.e., the EPC network [21]. Non-3GPP access is the general name used to denote any radio access that is not defined by 3GPP, such as WLAN, Wi-Max, and CDMA2000 packet data radio access. The principles of the interworking architecture that were concluded by the SA working group of 3GPP were that a) the interworking between 3GPP and non-3GPP networks to support terminal mobility occurs at the network layer and b) IETF-based mobility protocols should be used as a basis for the design of solutions to mobility between accessing the EPC via non-3GPP access and via 3GPP access.

Related functions were introduced into the specifications step by step. In Rel-8, the interworking architecture was defined, and the assumption has been made that the terminal can only access the EPC through either one 3GPP access or one non-3GPP access at the same time; handover procedures for PDN connections between different accesses are also defined, along with basic procedures for accessing the EPC via non-3GPP access such as attach, detach, PDN connection establishment, and disconnection.

The most important function that has been introduced since Rel-8 is called Access Network Discovery Selection Function (ANDSF). The ANDSF empowers the operator to control how the user discovers and selects a non-3GPP access network if several different non-3GPP access networks are available. In the Rel-9 timeframe, based on the assumption that the terminal can access the EPC via one 3GPP access and one non-3GPP access simultaneously, studies focused on the support of multiple PDN connections to the same APN, and studies of IP flow mobility using host-based mobility protocols such as DSMIP have been concluded, but due to time limitations, the related standardization works were conducted in Rel-10. In Rel-11, IP flow mobility using network-based mobility (NBM) protocols, i.e., PMIP and GTP, was studied, but the studies were not concluded; related works continued within Rel-12 and Rel-13, e.g., how to support GTP on S2a.

(1) Access Network Discovery and Selection

The ANDSF function uses an application layer method (e.g., the open mobile alliance-device management) to provide operator-defined policies to the UE and non-3GPP access network discovery and selection information in the UE's vicinity, based on the UE-provided location, which then will be used by the UE and the user to select the non-3GPP access networks.

These policies consist of inter-system mobility policy (ISMP) and inter-system routing policy (ISRP). The ISMPs are used to select one non-3GPP access when the UE can only operate one access at any one time. For UE that can operate one 3GPP access and one non-3GPP access at the same time, the ISRPs are used for routing of the different PDN connections, where the IP flows over one 3GPP access or a particular type of non-3GPP access. Both the ISMPs and the ISRPs are operator-defined and

only influence the non-3GPP access selection; they do not affect 3GPP access selections.

(2) Architecture

We call the interworking architecture EPC-centric, because all non-3GPP accesses are integrated in this architecture, even for the 3GPP2-defined High-Rate Packet Data (HRPD) access. The EPS is the central control point for all functions related to the interworking procedures, such as AAA, mobility management (including handover between accesses), QoS, policy control, and charging. An operator-deployed I-WLAN can also be migrated to networks based on this architecture.

Within the EPC-centric architecture, both network-based mobility and host-based mobility management mechanisms are supported. For HBM, DSMIP, or MIP client software is explicitly needed in the terminal to exchange mobility-related signaling with the network. For NBM, mobility-related signaling is exchanged between the network nodes to maintain IP session continuity for terminal mobility. For both HBM and NBM, the PDN-GW is the user plane anchor for mobility between 3GPP access and non-3GPP access (whether there is another entity that acts in the role of a PDN GW that is located outside the 3GPP network is beyond this scope of this book, and also beyond the scope of 3GPP). Two types of non-3GPP access networks are defined: trusted non-3GPP access networks and untrusted non-3GPP access networks. Whether each non-3GPP access network is trusted or untrusted is mainly decided on the basis of the operator policy, and not on the security characteristics of the access technology itself. For example, those Wi-Fi networks deployed by operator itself are considered as trusted. For a terminal accessing EPC via the untrusted non-3GPP access network, additional security methods, e.g., a security tunnel between the terminal and the evolved PDG (ePDG), are needed to secure the PLMN. Two non-roaming network architectures are depicted in Fig. 5.39. Roaming network architectures have also been defined, and details can be found in TS23.402 [21].

In (a), for trusted non-3GPP access, the reference point S2a provides a user plane with related controls and mobility support between the trusted non-3GPP IP Access network and the gateway, and for the untrusted non-3GPP access, the reference point S2b provides the user plane with the related control and mobility support between the evolved PDG and the PDN GW, and a security tunnel between the terminal and the ePDG is needed. If PMIPv6 is used over S2a; then, the PDN GW is the Local Mobility Anchor (LMA) and the Mobile Access Gateway (MAG) is located in the trusted non-3GPP access network. If the MIPv4 FA mode is used over S2a; then the home agent function is located within the PDN GW, and the foreign agent is located within the trusted non-3GPP access network. If PMIPv6 is used over S2b, the PDN GW is the LMA and the MAG is located within the ePDG. Technically, the MIPv4 FA mode may also be used over S2b; however, it has not been standardized by 3GPP because of a lack of requirement.

In (b), DSMIPv6 is used, and S2c is the reference point that provides the user plane with the related control and mobility support between the UE and the

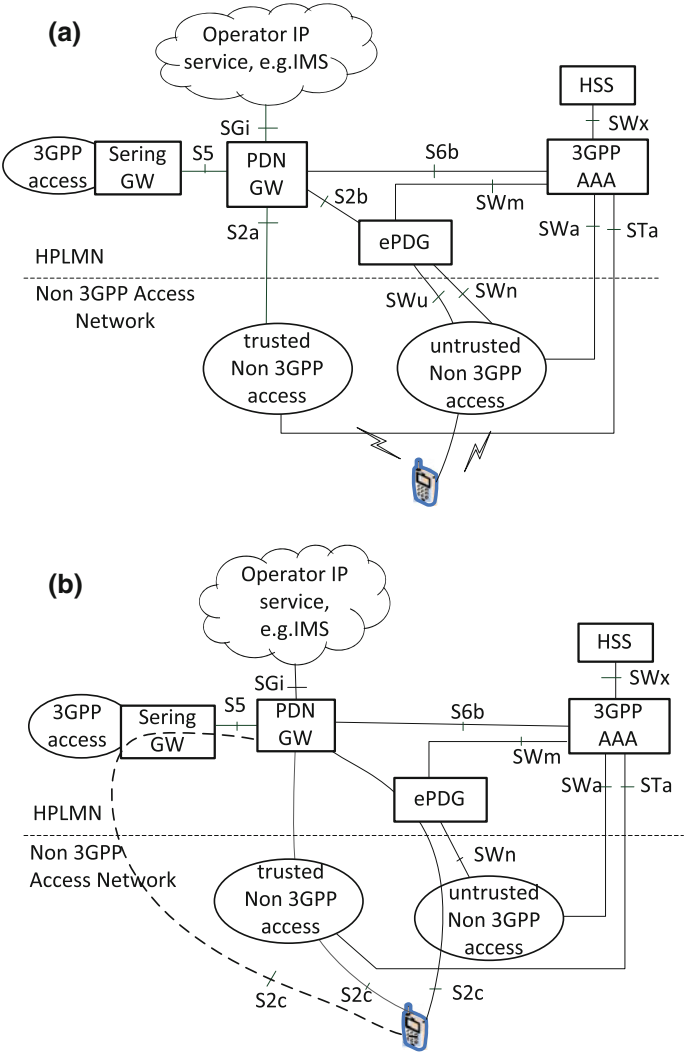


Fig. 5.39 Non-roaming architectures by reference points: **a** S2a and S2b. **b** S2c. Reprinted from 3GPP TS23.402 [21]

PDN-GW. The S2c reference point can be implemented in three cases: trusted non-3GPP access, untrusted non-3GPP access, and 3GPP access. For untrusted non-3GPP access, a security tunnel is needed between the terminal and the ePDG. For the above cases, the HA functionality is collocated with the PDN-GW.

(3) IP Mobility Management Selection and IP Session Continuity

To support mobility management properly, and to maintain IP session continuity, the network and the terminal should use the same mobility management

mechanisms. The operator may statically configure the use of a single IP mobility management mechanism on the network side, and let the terminal know the mechanism that is used by the configuration. Alternatively, the operators may also deploy multiple IP mobility mechanisms in their networks. In this case, the IP mobility management selection (IPMS) is needed to ensure that the terminal and the network use the same mechanism.

The IPMS is not needed for a terminal that attaches through 3GPP access, e.g., power-on attachments, to the EPC, because the NBM is always used in this case. The IPMS procedure is performed before an IP address can be allocated to the terminal in following cases: initial attachment via non-3GPP access, and non-optimized handover from 3GPP access to non-3GPP access.

As part of the authentication signaling exchange procedure in the above cases, the HSS and the AAA decide which mobility management mechanism is to be used, based on the terminal's capabilities, the local access network capabilities and the home network capabilities and operator policies. If the terminal has indicated its supported mobility mechanisms, the network will provide the selected mobility mechanisms to the terminal. If the network is not aware of the terminal's capabilities, i.e., the terminal did not explicitly indicate its capabilities, and then the NBM is used. During this authentication procedure, the HSS and the AAA send an indication that indicates one of following decisions to the non-3GPP access network: (a) allocate a local IP address to the terminal, which means that DSMIP will be used; (b) allocate a MIPv4 foreign agent address to the terminal, which means that MIPv4 FA mode will be used; or (c) use the NBM mechanism. For (c), the HSS/AAA (and IPMS) does not make the selection between GTP and PMIP. The operator may configure the deployment of either GTPv2 or PMIPv6 on the S5/S8 interfaces and at the S2b reference point, and the terminal is not aware of which NMB protocol is used within the network.

To support seamless mobility between the 3GPP access and the non-3GPP access, the IP address preservation is used to maintain IP session continuity. During interaccess mobility, the PDN GW selected within the new access is the same one that was used in the old access, and the IP addresses that have been allocated to the terminal in the new access are the same as the addresses used in the old access.

For NBM cases, the terminal indicates its capability of supporting IP address preservation and the IP addresses (IPv4 address and/or IPv6 prefix) allocated when it attached via the old access (e.g., 3GPP access) to the network during the inter-access handover procedures (e.g., handover to non-3GPP access). This will allow the PDG to reallocate the same IP addresses to the terminal when the terminal is connected via the new access (e.g., non-3GPP access).

If the terminal handover is from 3GPP access to non-3GPP access and the HBM is used, the non-3GPP access network first allocates a new IP address to the terminal, which can be used as a local IP address or as the FA Care-of Address (CoA) IP address, depending on whether DSMIPv6 or MIPv4 is used, respectively. Then, IP address preservation is achieved by performing the DSMIPv6 procedure over S2c or the MIPv6 FA CoA procedure over S2a. If the terminal handover is from a non-3GPP access where HBM is used to 3GPP access, the terminal knows

that it is returning to its home link and deregisters the binding association between the local address or the FA CoA and its home address via DSMIPv6 or MIPv4 procedures, respectively. In this case, the 3GPP access is configured as the home link. However, non-3GPP access can also be configured as a home link, which is not preferred by the 3GPP.

(4) GTP support over S2a in Rel-12

To support GTP over S2a, the following functions are needed for a Trusted WLAN (TWAN), where one or more APs connect to the TWAN Gateway (TWAG), and the TWAG terminates the S2a interface to the PDN GW. The UE-TWAG protocol (i.e., the WLAN control protocol) has been introduced to control (e.g., setup and teardown) PDN connections over a TWAN. There is a per-UE point-to-point link between the UE and the TWAG, and the TWAG maps the p2p link to and from the GTP tunnels. Since such limitations in Rel-11, i.e., no UE and WLAN impacts were allowed are removed since Rel-12 and onwards, necessary enhancements to the UE and the TWAN are defined, multiple simultaneous PDN connections over TWAN are supported in Rel-12.

(5) Future work

Interworking between 3GPP systems and other wireless systems has been studied for many years [21]. Numerous technologies related to mobility management from the IETF have been incorporated into 3GPP specifications related to such interworking. The I-WLAN is considered to be an “overlay” solution, which means that it has no impact on WLAN networks. However, because few terminals support the required client software, e.g., DSMIP, there is limited deployment of I-WLAN to date. With the increased numbers of WLAN-enabled terminals that support USIM/SIM-based AAA, the interworking architecture based on EPS become feasible for deployment.

In Rel-12 and Rel-13, new study areas include: interworking of Wi-Fi and LTE networks from the radio perspective, and how to perform WLAN access network selection for such cases, autonomous authentication and authorization based on policies from WLAN based on HotSpot2.0 (specifications developed by the Wi-Fi Alliance to support global Wi-Fi roaming, which focuses on network selection, security, and policy provision aspects) and those policies from 3GPP operators, network-based IP flow mobility, and so on.

References

1. Iqbal MK, Iqbal MB, Rasheed I et al (2012) 4G Evolution and Multiplexing Techniques with solution to implementation challenges. 2012 international conference on cyber-enabled distributed computing and knowledge discovery (CyberC), pp 485–488. doi:[10.1109/CyberC.2012.88](https://doi.org/10.1109/CyberC.2012.88)
2. Jamalipour A, Wada T, Yamazato T (2005) A tutorial on multiple access technologies for beyond 3G mobile networks. IEEE Commun Mag 43(2):110–117. doi:[10.1109/MCOM.2005.1391509](https://doi.org/10.1109/MCOM.2005.1391509)

3. Jain RK, Katiyar S, Agrawal NK (2011) Hierarchical cellular structures in high-capacity cellular communication systems. (IJACSA) Int J Adv Comput Sci Appl 2(9):51–57 arXiv: 1110.2627
4. <http://www.smallcellforum.org/aboutsmallcells-small-cells-what-is-a-small-cell>. Accessed 23 Mar 2013
5. Nakamura T, Nagata S, Benjebbour A et al (2013) Trends in small cell enhancements in LTE advanced. IEEE Commun Mag 51(2):98–105. doi:10.1109/MCOM.2013.6461192
6. <http://standards.ieee.org/about/get/802/802.11.html>. Accessed Jan 2014
7. <https://www.wi-fi.org/>. Accessed July 2014
8. Chini P, Giambene G, Kota S (2010) A survey on mobile satellite systems. Int J Satell Commun Netw 28(1):29–57. doi:10.1002/sat.941
9. Elbert BR (2008) Introduction to satellite communication, 3rd edn. Artech House, Norwood
10. Fan Linghang, Cruickshank Haitham, Sun Zhili (2007) IP networking over next-generation satellite systems. Springer, Guildford
11. NMT DOC 450-1, Nordic Mobile Telephone-SYSTEM DESCRIPTION, August 1997
12. Chen S, Wang Y, Ma W et al (2012) Technical innovations promoting standard evolution: from TD-SCDMA to TD-LTE and beyond. IEEE Wirel Commun 19(1):60–66. doi:10.1109/MWC.2012.6155877
13. GPP TS 25.420 UTRAN Iur interface general aspects and principles
14. <http://www.3gpp.org/specifications/67-releases>. Accessed Jan 2014
15. GPP TS 23.401 “General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access” v. 12.2.0
16. GPP TS 23.060 “General Packet Radio Service (GPRS); Service description; Stage 2” v. 12.2.0
17. GPP TS 23.272 “Circuit Switched (CS) fallback in Evolved Packet System (EPS); Stage 2” v. 12.0.0
18. [LTE book] Magnus Olsson, Shabnam Sultana, et al (2013) “EPC and 4G Packet Networks (second edition)”
19. GPP TS 23.122: “Non-Access-Stratum (NAS) functions related to Mobile Station (MS) in idle mode”
20. <http://www.3gpp.org/>. Accessed Jan 2015
21. GPP TS 23.402, “Architecture enhancements for non-3GPP accesses,” V12.0.0
22. GPP2 P.S0001-B, Wireless IP Network Standard, Version 1.0.0, October 25, 2002
23. GPP2 X.S0013-000-A, All-IP Core Network Multimedia Domain Overview, Version 1.0, November, 2005
24. GPP TS 36.300 “Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN)”; Overall description; Stage 2; Version. 11.7.0
25. GPP TS 43.051 “Overall description; Stage 2”
26. GPP TS 25.401 “UTRAN overall description”
27. GPP TS 25.410 “UTRAN Iu interface: General aspects and principles”
28. GPP TS 24.301 “Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3” v. 12.2.0
29. GPP TS 24.008 “Mobile Radio Interface Layer 3 specification; Core Network Protocols; Stage 3”
30. GPP TS 29.281 “General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U)” v. 11.6.0
31. GPP TS 29.060 “General Packet Radio Service (GPRS); GPRS Tunneling Protocol (GTP) across the Gn and Gp interface” v. 12.2.0
32. GPP TS 31.102 “Characteristics of the Universal Subscriber Identity Module (USIM) application”
33. GPP TS 36.304 “Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) procedures in idle mode” v. 11.5.0
34. GPP TS 21.905 Vocabulary for 3GPP Specifications

35. Labedz G, Felix K, Lev V, Schaeffer D (1987) Handover control issues in very high capacity cellular systems using small cells. In: International conference on digital land mobile radio communication
36. Nanda S (1993) Teletraffic models for urban and suburban microcells: cell sizes and handoff rates. *IEEE Trans Veh Technol* 42(4):673–682
37. Stüber GL (2011) Principles of mobile communication. Springer, Berlin
38. GPP TS 36.331 Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification v. 11.5.0
39. Ekiz N, Salih T, Kucukoner S et al (2005) An overview of handoff techniques in cellular networks. *Int J Inf Technol* 2(3):132–136
40. Aggarwal S, Tomar S K, Aggarwal A et al (2013) Survey and taxonomy of soft handoff in 3rd generation CDMA based System
41. Li B, Xie D, Cheng S et al (2005) Recent advances on TD-SCDMA in China. *IEEE Commun Mag* 43(1):30–37. doi:[10.1109/MCOM.2005.1381872](https://doi.org/10.1109/MCOM.2005.1381872)
42. GPP TS 43.129 “Packet-switched handover for GERAN A/Gb mode; Stage 2”
43. Bikos A, Sklavos N (2013) LTE/SAE security issues on 4G wireless networks. *IEEE Secur Priv* 11(2):55–62. doi:[10.1109/MSP.2012.136](https://doi.org/10.1109/MSP.2012.136)
44. GPP TS 33.102 “3GPP Security: Security architecture” Version 12.0.0
45. GPP TS 33.401 “3GPP System Architecture Evolution (SAE): Security architecture” V12.9.0
46. GPP TS 33.402 “3GPP System Architecture Evolution (SAE); Security aspects of non-3GPP accesses”
47. ETSI (2001) TR 101 957 “Requirements and architectures for interworking between HIPERLAN/2 and 3rd generation cellular systems”
48. GPP TR 22.934 “Feasibility study on 3GPP system to Wireless Local Area Network (WLAN) interworking” v. 11.0.0
49. GPP TS 23.234 “3GPP system to Wireless Local Area Network (WLAN) interworking; System description” v. 11.0.0
50. GPP TS 23.216: “Single Radio Voice Call Continuity (SRVCC); Stage 2”
51. GPP TS 23.327 “Mobility between 3GPP-wireless local area network (WLAN) interworking and 3GPP systems” v. 11.0.0

Chapter 6

Mobility Management at Network Layer

This chapter focuses on basic concept and method of mobility management solutions at the network layer. Typical protocols such as Mobile IP (MIP), Dual Stack Mobile IP, and Proxy Mobile IP are introduced, as well as related key technologies. The support of IP flow mobility, network mobility solutions, and security mechanism related to mobility management at network layer are also briefly described.

6.1 Overview

At early age of Internet, hosts (e.g., mainframe computers, desktop computers, printers) are connected to the network via Internet Protocol (IP) [1]-based wired links, e.g., via Ethernet, cable, and xDSL. Due to the development of wireless communication, more and more hosts (e.g., laptop, tablet) require communication during or after movement. Lots of protocols of different layers and solutions which support host mobility appeared. Network-layer mobility management technology based on IP protocol, is featured by its agnostic to the underlying access technologies. It is also transparent to the upper-layer protocols and applications, and has become the focus of mobility management research.

The Internet is designed for fixed or wired networks, where each host is assigned one IP address or more IP addresses (e.g., router) via either static configurations or dynamic interaction with network (e.g., via Dynamic Host Configuration Protocol (DHCP) [2]). However, exactly what IP address could be assigned to a host is based on network topology, rather than based on host's identity. Actually, an IP address actually identifies a network attachment point or an interface to the network, rather than the host itself. The host has fixed network attachment point, and the corresponding IP address could be considered as this host's identity as well.

If a host changes its network attachment point, it will be assigned a new IP address. Thus, ongoing communication based on IP would be interrupted due to the following reasons: (a) IP packets destined to the host's old IP address arrived at old

network would not be forwarded to new attachment point and (b) changing IP address would result in rebinding or reregistration of new IP address to upper-layer application in most cases, and the corresponding procedure incurs long delay or interruption which deteriorates user experience.

Some solutions have been proposed to resolve these problems. One of them is to update new IP address with corresponding part as soon as possible. However, this solution does not hide host mobility to upper-layer protocols and applications. For example, all application that is based on Transmission Control Protocol (TCP) [3] that requires fixed IP address during the lifetime of the TCP connection could not guarantee service continuity. Another proposal is to deploy host-specific routing within whole IP network. In this solution, the IP address allocated by host's home network would keep unchanged even the host has moved to another network attachment point. All routers within the network update their routing table with this host address, that is, a host-specific routing entry is created within the whole network. The signaling cost for host-specific route is high for a host. Considering that more and more hosts need mobility support, this solution would introduce huge routing update signaling. That is, this solution has poor scalability and efficiency.

A network-layer mobility management solution could shield the difference of underlying access technologies to upper-layer protocols and applications. Internet Engineering Task Force (IETF) had worked out a solution in 1990s, which is named as MIP [4–8]. It hides mobility of host from its correspondent communication node by registering mobile node's (MN's) care-of address (CoA) to the home agent (HA); thus, this mobile host can continue using its home address even if it leaves its home link. In this case, packets that target at host's home address would be tunneled or redirected to this CoA. This mechanism requires the MN to detect mobility, register current CoA, and maintain the registration periodically, so it is called the host-based mobility management solutions.

IETF also developed a network-based mobility management scheme, Proxy Mobile IPv6 (PMIPv6) [9], where it is the network that is responsible for host mobility detection and registration without the involvement of the host.

Besides the host-based and network-based mobility management solutions it only handles host mobility, a solution that handles network mobility named as NEMO (Network MObility) [10] is proposed by IETF. It would be introduced later in this chapter, along with the IP flow mobility [11], a fine granularity mobility solution that is capable of moving IP flows between different interfaces.

6.2 Mobility Management Solutions at Network Layer

6.2.1 *Host-Based Mobility Management for Host Mobility*

The research related to management technology at network-layer mobility began in the early 1990s [12]. In 1991, Columbia MIP [13] proposed to use the virtual mobile subnet and IP in an IP tunnel encapsulation to support the host mobility.

Almost at the same time, the Virtual Internet Protocol (VIP) [14] proposed to use a special router to record the location of the MN and defines new IP header options for data transmission. A MN protocol using the loose source routing of the IP protocol [15] was proposed. In the year 1994, based on analyzing the design ideas and implementation methods of the above three solutions, Myles and Perkins proposed a new node mobility support solution named MIP, and this solution was then developed by the IETF working group afterward and became into the MIP protocol.

(1) Mobile IPv4

The MIP protocol provides mobility support for MN in the Internet, ensuring that MN's ongoing communication session is not interrupted when an MN changes its network attachment point. The first version was published in 1996, that is Request For Comments (RFC) 2002 [4], and an updated version published in 2002, that is RFC 3344 [5]. In 2010, the RFC 3344 [6] was obsolete by RFC 5944 [6].

The MIPv4 protocol defines the following functional entities and concepts:

- **Mobile Node (MN)**
MN refers to a host or router that may change its point of attachment from one network to another.
- **Home Agent (HA)**
HA refers to a router on an MN's home link that tunnels datagrams to the MN when this MN is away from home link, if the HA knows the binding of CoA and home address of this MN.
- **Foreign Agent (FA)**
FA is a router on a foreign network that provides routing service to the MN temporary resides on this network. The FA detunnels and delivers the datagrams that are tunneled by the MN's HA. For datagrams sent by the MN, the FA may serve as a default router for registered MNs.
- **Home Network**
Home network is a network whose network prefix matches that of an MN's home address.
- **Foreign Network**
Foreign network is any network other than the MN's home network.
- **Home Address**
Home address refers to an IP address that is assigned for an extended period of time to an MN. It remains unchanged regardless of where the node is attached to the Internet.
- **Care-of Address**
The CoA is the address which is used by the MN to receive or send datagrams when they are in a foreign network. There are two types of care-of address: A "foreign agent care-of address" is an address of a FA with which the MN is registered, and a "colocated care-of address" is an externally obtained local address that the MN has associated with one of its own network interfaces.

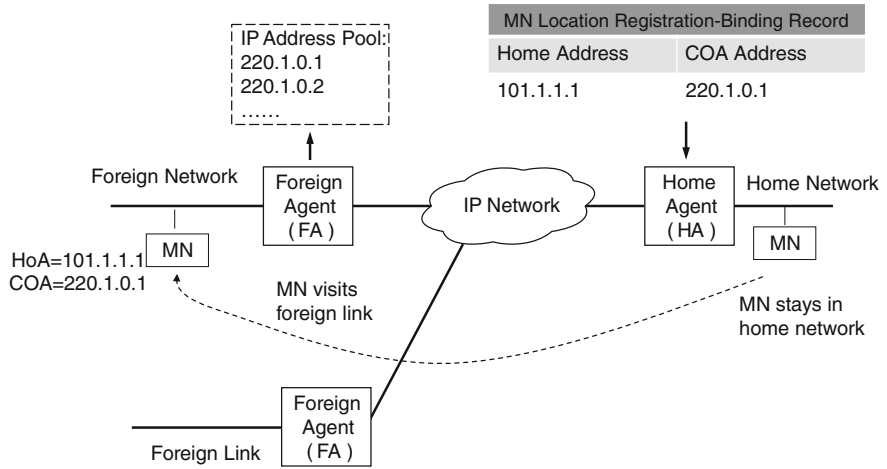


Fig. 6.1 An example of network deployment for MIPv4

- **Correspondent Node (CN)**
The peer node communicating with the MN is called a CN. A CN may locate at anywhere in the network.

An example of the network deployment using MIPv4 is shown in Fig. 6.1. As illustrated in Fig. 6.1, the MN receives its home address (101.1.1.1) in the home link. When it moves to a foreign link, it acquires a CoA (220.1.0.1) from the FA. The MN needs to register its CoA to the HA by sending a location registration request message. The HA updates and stores the binding between its CoA and home address and returns a registration reply message. After that, when packets sent by a CN to the MN reach the MN's home network by normal IP routing (i.e., using MN's home address), these packets are intercepted by the MN's HA and then tunneled to the foreign network where the MN is currently attached by using the care-of address.

The above process is shown in Fig. 6.2. In the step of agent discovery, the MN detects that it is on a foreign link and obtains a CoA. The MN registers its CoA in the HA in the step of location registration. With successful registration, the HA intercepts, encapsulates, and tunnels the packets to the MN.

It should be noted that although an MN can only receive packets destined to its home address via the HA, it can send packets directly to the CN using its CoA. This is known as the triangle routing problem, which is shown in Fig. 6.3. Depending on the location of the MN and the CN, there is a chance that the routing path of the outgoing packets (see the dotted line in Fig. 6.3) is much shorter than that of the incoming packets (see the solid line in Fig. 6.3). This problem is caused by the packet to MN which needs to be intercepted by HA and forwarded to the MN, thus increase the signaling load of HA and decrease packet forwarding efficiency, especially in the case where MN is located in a foreign network far away from the home network, and the CN also is located in this foreign network.

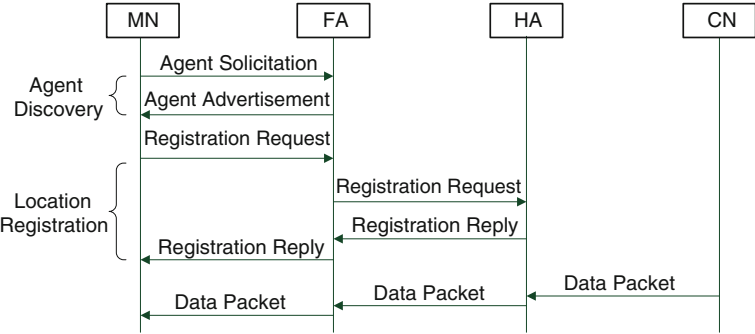


Fig. 6.2 Location registration and data forwarding in MIPv4

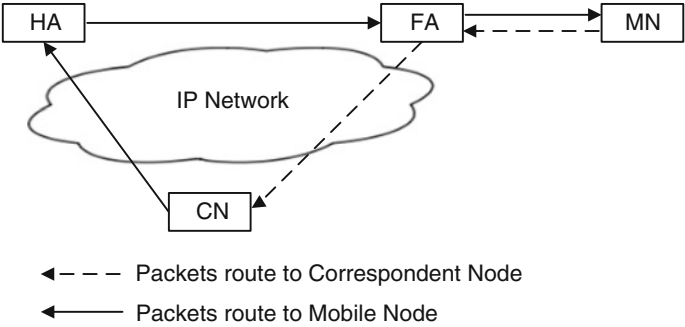


Fig. 6.3 Triangle routing problem in MIPv4

(2) Mobile IPv6

In 2006, IETF published the first Mobile IPv6 (MIPv6) protocol standard—RFC 3775 [7]—and then updated it with RFC 6275 [8] in 2011. Compared with Mobile IPv4 (MIPv4), the functionality and performance have been improved greatly.

Benefiting from the built-in characteristics of IPv6 such as large address space, address autoconfiguration, and neighbor discovery mechanism, MIPv6 is much concise and simpler. Although MIPv6 shares some basic ideas of MIPv4, there are many important differences from MIPv4 after rethinking:

- There is no need to deploy special routers as FAs, since the MN uses the IPv6 neighbor discovery mechanism or address autoconfiguration access to obtain its CoA.
- Besides HA that can establish binding between home address and care-of address of the MN, the corresponding node can also establish such binding; thus, there is no triangle routing problem in MIPv6.

- MIPv6 provides a dynamic HA discovery mechanism, which allows the MN to discover the new HA and register its CoA when HA reconfiguration or replacement happens.
- MIPv6 offers a new security mechanism, including IPv6 Neighbor Unreachability Detection and ingress filtering.

The functional entities in MIPv6 are the same as that in MIPv4 except that FA is not needed in MIPv6. One example of network deployment for MIPv6 and MIPv6 registration procedure is shown in Figs. 6.4 and 6.5, respectively. In this deployment, the MN receives an IPv6 address 3FFE:0:0:7::1/64 as the home address. When it moves to the foreign network, an IPv6 CoA 3FFE:0:0:9::1/64 is allocated.

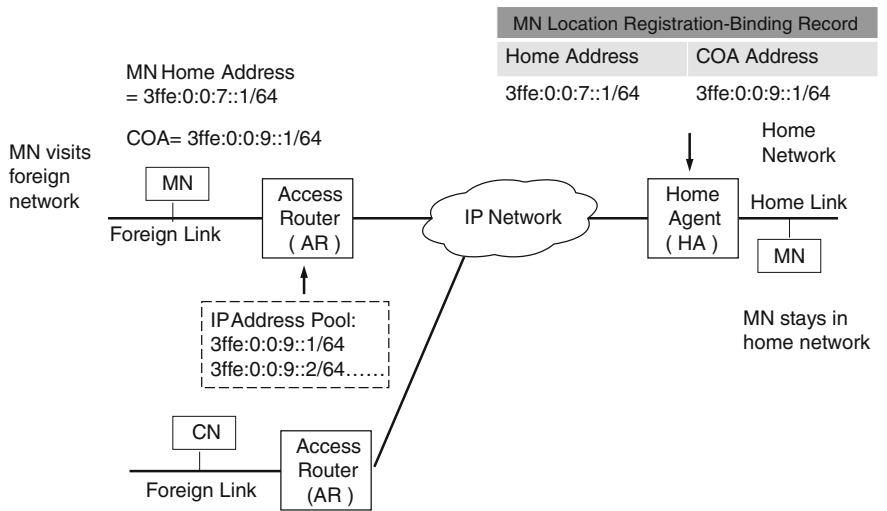


Fig. 6.4 An example of network deployment for MIPv6

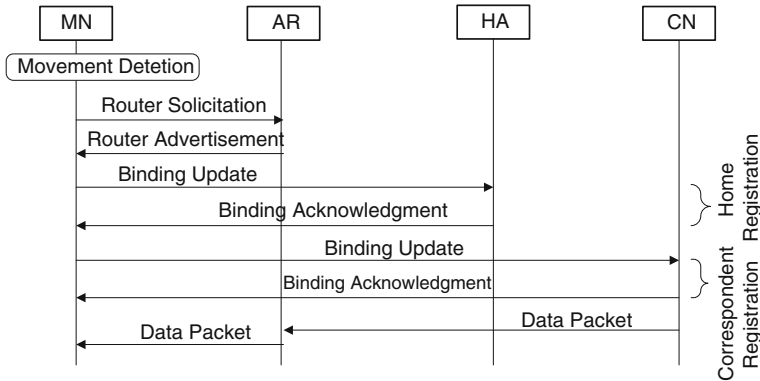


Fig. 6.5 Location registration and data forwarding in MIPv6

The MN first uses the IPv6 neighbor discovery mechanism to perform mobility detection, i.e., the MN determines whether it has moved to a foreign link by examining the subnet prefix included in the Router Advertisement (RA), or by detecting the unreachability of the default router. When the MN moves into a foreign network, it acquires its CoA through an IPv6 neighbor discovery mechanism such as stateless autoconfiguration and then sends a Binding Update (BU) message to its HA for registering its CoA. The HA then responds with a binding acknowledgment message to the MN.

The BU can be sent to any CN, as well. Upon the receipt of such a BU message, the CN retrieves the home address and CoA pairs it maintained and updates the binding accordingly.

After successful registration, the MN can communicate with the correspondent node by using its home address or care-of address. There are two modes for communication between the MN and a CN: a bidirectional tunneling mode and a route optimization mode, which are based on the HA registration and HA + CN registration, respectively.

A. Bidirectional tunneling mode

In this mode, communication between the MN and correspondent node must traverse the HA. Packets from the CN are routed to the HA and are then tunneled to the MN (see the solid line shown in Fig. 6.6). Packets to the CN are first tunneled from the MN to the HA (referred to as “reverse tunneled”) and are then routed normally from the home network to the CN (see the dotted line in Fig. 6.6).

This mode does not require correspondent node supporting MIPv6, and correspondent registration is also not performed.

B. Routing optimization

This mode requires a CN supporting MIPv6, i.e., correspondent registration should be performed before communication between CN and MN.

Before sending a packet, the CN examines its binding cache to check whether any entries for the destination address (i.e., care-of address) of such packet exist. If a binding entry exists, then it sends this packet to the MN using CoA (see the solid line in Fig. 6.7). Instead, the CN sends the packet to the MN using the home address as the destination address (see the dotted line in Fig. 6.7).

For the MN-to-CN direction, packets are sent directly to the CN from the MN.

Fig. 6.6 Bidirectional tunneling mode in MIPv6

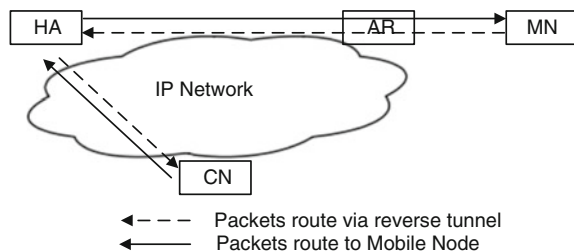
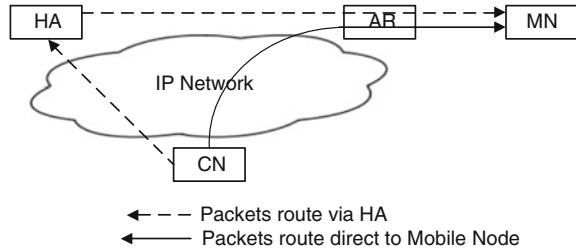


Fig. 6.7 Route optimization mode in MIPv6



Within route optimization mode, triangle routing problem does not exist due to the CN registration. If the MN receives a packet destined to its CoA and forwards it to the upper layer, the packet dropped since the existing upper-layer communication is bound to the home address. In order to resolve this, a new IPv6 destination option—the home address option—has been introduced. With this option, the service continuity is achieved by hiding network-layer mobility event to upper layers.

When the MN is away from home network and sends packet to the CN, it uses the home address option with its home address while using its CoA as the source address of the outgoing IPv6 packet. When a CN receives such packet with home address option included, it swaps the value of the home address option and IPv6 source address. Similarly, when a CN sends packet to the MN, it first checks in its binding cache whether the entry of the destination address (the MN's home address) of the packet exists. If so, then the CN sends the packet by using the CoA as the destination address, and it includes the home address in the type 2 routing header. When the MN receives such a packet, it swaps the destination address with the home contained in the type 2 routing header before passing such a packet to the upper layer. In this way, the network-layer mobility of MN is transparent to the upper layer.

In summary, the MIPv6 takes advantages of the neighbor discovery mechanism, an optimized routing header, and some extension options of IPv6. In MIPv6, performance and efficiency are improved compared with those of MIPv4. Table 6.1 provides a comparison between MIPv4 and MIPv6.

(3) Dual Stack Mobile IPv6 (DSMIPv6)

Although MIPv6 inherited several advantages from IPv6 such as large address space, the actual transition from IPv4 to IPv6 will take a considerably long time. So it is not likely that MNs will only use IPv6 addresses at the beginning of IPv6 deployment, since many applications still use IPv4. It is also reasonable to consider the issue that how an IPv6-capable MN maintains the ongoing session when moving to an IPv4 network. To solve these problems, IETF started the study on Dual Stack Mobile IP and released RFC 5555 [16] in 2009.

The DSMIPv6 reuses MIPv6 to perform binding registration. To enable the use of IPv4 home address or CoA, MIPv6 is extended with a couple of new mobile headers. Besides, to ensure that DSMIPv6 messages can be sent both in IPv4 and in IPv6 networks, IP-in-IP encapsulation is used for MN and HA in certain cases.

Table 6.1 Comparison between MIPv4 and MIPv6 protocols

Function	MIPv4 (RFC 5944)	MIPv6 (RFC 6275)
Functional entities	HA, FA, and MN	HA, MN, and CN
Mobility detection	FA discovery	IPv6 neighbor discovery
CoA configuration method	DHCP or manual configuration	Stateless autoconfiguration, DHCP, or manual configuration
Length of address	32 bit	128 bit
Type of CoA	Configured CoA or FA CoA	Configured CoA
Data forwarding method	Tunneling	Tunneling or routing optimization
Tunneling mode	HA-FA, decapsulation of foreign agent; HA-MN, decapsulation of mobile node	HA-MN, decapsulation of mobile node
Dynamic home agent address discovery (DHAAD)	Directed broadcast	Anycast addressing
Route optimization	None, cooperates with other protocols	Integrated
Security considerations	None, cooperates with other protocols	Built-in IPSec

It is assumed that the MA and HA support both IPv4 and IPv6. However, the CN is not required to support DSMIPv6 and is kept transparent whether DSMIPv6 is used. The MN needs to know the IPv4 address of the HA besides its IPv6 address, to enable IPv4 routing.

The application of DSMIPv6 can be discussed in the following three scenarios:

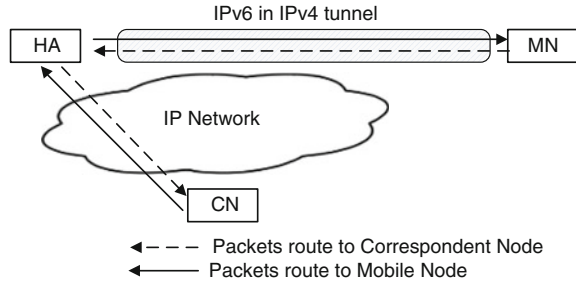
A. Scenario1: IPv4-only foreign network

In this scenario, the MN will be assigned with a globally unique IPv4 CoA. The mobile mode should register this CoA to its HA using DSMIPv6 signaling. If the MN has an IPv4 home address and an IPv6 home address, each home address needs to be bound with the CoA so that both IPv4 and IPv6 applications can continue with the MN’s home address.

The BU message is extended with an IPv4 care-of address option to carry the IPv4 CoA. If the MN receives an IPv4 home address, the BU message should also include an IPv4 home address option to carry the IPv4 home address to the HA, besides the already included IPv6 home address.

In order to transfer the BU message to the HA via the IPv4 foreign network, the MN shall encapsulate the message into an IPv4 packet. The source address of this IPv4 packet is the CoA of the MN, and the destination address of this IPv4 packet is the IPv4 address of the HA. When receiving and accepting the BU message, the HA creates a new entry in its binding cache for the MN’s IPv6 home address. Since the IPv4 home address option is also received, the HA creates another entry for the MN’s IPv4 home address. The HA returns a Binding Acknowledgement to the MN.

Fig. 6.8 IPv6 in IPv4 tunneling in case of IPv4-only foreign network



Hence, all packets destined to the MN's home IPv4 or IPv6 address will be tunneled by the HA, which encapsulates these packets in IPv4, where the source address is the HA's home IPv4 address and the destination address is the MN's IPv4 care-of address. The tunneling model is illustrated in Fig. 6.8.

B. Scenario2: IPv4-only private network

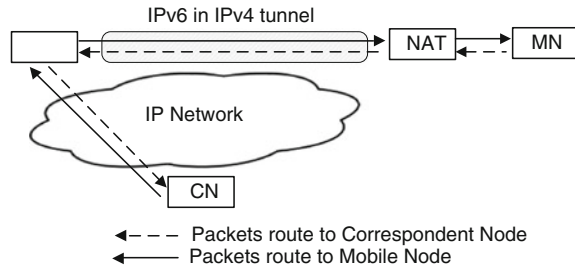
When the MN moves to an IPv4 private network which connects to the Internet via a NAT (network address translation), it will be assigned with a private IPv4 CoA. To register the IPv4 CoA in the HA, the MN sends a BU message destined to the HA's IPv4 address. Although in this scenario the HA is reachable via its IPv4 address, the HA cannot send Binding Acknowledgement message to the MN by using the IPv4 CoA since it is not routable. To solve this problem, the MN needs to encapsulate the BU message into a UDP (User Datagram Protocol)/IPv4 datagram, in which the IPv4 source address is the IPv4 CoA and IPv4 destination address is the IPv4 address of the HA. In the DSMIPv6 BU message, the MN includes the IPv4 home address option and IPv4 care-of address option.

Upon receiving the BU message contained in the UDP/IPv4 datagram, the HA compares the IPv4 address contained in the IPv4 care-of address option with the source IP address of the IPv4 outer header. If they are different, it means that NAT exists between the MN and the HA, which should be traversed. The HA encapsulates the Binding Acknowledgement in UDP and then in IPv4, with the HA's IPv4 addresses in the source IP address field and the source IPv4 address contained in the received IPv4 packet in the destination IP address field. In the inner IPv6 BU message, the HA also includes a NAT detection option, which indicates to the MN the existence of NAT and that UDP encapsulation is required.

After the tunnel between the HA and MN is established, communication between the MN and the CN is tunneled via the HA, as shown in Fig. 6.9.

- When the HA receives packets destined to the MN's home address, it encapsulates each packet in UDP and then IPv4, where the source IP address field is filled with the HA's IPv4 address and the target IP address field is filled with the source IPv4 address received in the IP packet carrying the BU.
- When the MN sends IPv6 data to the CN, it first encapsulates the data in UDP and IPv4 and then tunnels it to the HA. The source address of the IPv4 outer

Fig. 6.9 IPv6 in IPv4 tunneling in case of IPv4-only private network



header is the IPv4 CoA, which will be replaced with a routable IPv4 address when traversing the NAT. The target address of the IPv4 outer header is the HA's IPv4 address.

C. Scenario3: IPv6 foreign networks

When an MN moves into an IPv6 foreign network, it will get a globally unique IPv6 CoA. The MN sends a BU message to the HA's IPv6 address as defined in RFC 6275. The BU may also include the IPv4 home address option. When receiving the BU, the HA creates two binding cache entries, one for the MN's IPv6 home address and the other for the IPv4 home address.

Hence, when the HA receives a packet addressed to the MN's IPv4 home address or IPv6 home address, it tunnels the packet which is encapsulated in IPv6 to the MN's IPv6 CoA. When the MN sends communication data to CN, it also encapsulates and tunnels both IPv4 and IPv6 data in IPv6 to the HA.

6.2.2 Network-Based Mobility Management for Host Mobility

The host-based protocols could provide mobility management functions for hosts effectively, with the host participating in related signaling processes which require modification of the host's IP protocol stack. Meanwhile, if the MIP is working in a wide area network (WAN), when one MN moves within a relatively small area but far away from its home network, the massive BU signaling would lead to large delays. To provide better mobility management without MN's involvement, IETF proposed a network-based mobility management protocol in 2006: PMIPv6 (i.e., RFC5213) [9], to reduce signaling and handover delay. The PMIPv6 proposes a new functionality that resides on network side and provides mobility support to the MN, which moves within the PMIPv6 domain.

In PMIPv6, two new functional entities, Mobile Access Gateway (MAG) and Local Mobility Anchor (LMA), are introduced. When an MN moves into or within a PMIPv6 domain, MAG performs the signaling interaction with the LMA to

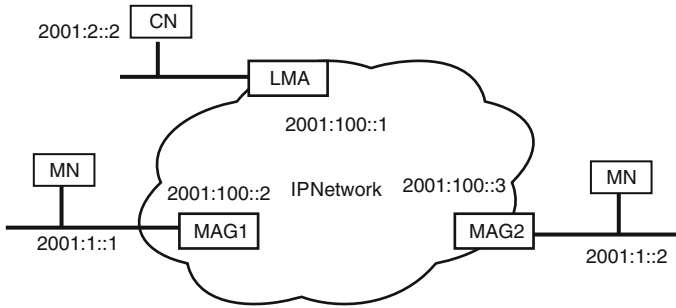


Fig. 6.10 An example of network deployment for PMIPv6

ensure that the MN could use the same IP address. A bidirectional tunnel is established between the LMA and the MAG to forward the packets to or from the MN. An example of network deployment for PMIPv6 is shown in Fig. 6.10.

As shown in Fig. 6.10, there are one LMA and two MAGs, MAG1 and MAG2, in a PMIPv6 domain. For visiting MNs, the MAG is responsible for establishing a bidirectional tunnel to the LMA. Bidirectional tunnel is used to route the data packets between the MAG and the LMA. The tunnel hides the topological architecture and enables the MN to use the address of its access link from the home network prefixes in the PMIPv6 domain. The tunnel is established dynamically, but it can be established in advance or could be removed if it is no longer required.

In Fig. 6.10, the CN is not on the same link as the MN; hence, the bidirectional tunnel between the MAG and LMA must be used to transfer the communication data between MN and CN. In case that an MN communicates with CN that is locally attached to the same access link, the MAG may optimize the packets delivery by routing packets locally without tunneling them to the LMA. However, in some systems, this may influence on the MN's account and policy enforcement, since the LMA is not in the traffic path and thus not able to apply any traffic policies or do any accounting for those traffics.

(1) Location Management in PMIPv6

In PMIPv6, the LMA is responsible for maintaining the MN's reachability state and is the topological anchor point for the MN's home network prefix (es). The MAG resides on the access link where the MN is located and performs the mobility management on behalf of the MN. There can be multiple LMAs in one PMIPv6 domain, and each serves a different group of MNs. Figure 6.11 shows the signal flow when the MN enters a PMIPv6 domain.

As shown in Fig. 6.11, when the MN attaches to an access link, the MAG sends a Proxy Binding Update (PBU) message to the MN's LMA. Upon accepting the PBU message, the LMA sends a Proxy Binding Acknowledgement (PBA) message, including the MN's home network prefix, and also creates a Binding Cache Entry (BCE) and sets up the endpoint of the bidirectional tunnel to the MAG.

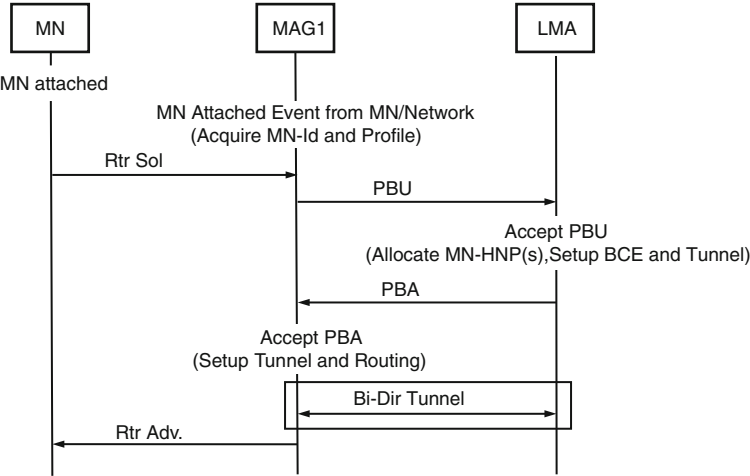


Fig. 6.11 MN attachment signaling flow in PMIPv6

Upon receiving the PBA message, the MAG sets up the endpoint of its bi-directional tunnel to the LMA and forwarding for the MN’s traffic. Then, it sends RA messages to the MN on the access link, advertising the MN’s home network prefix as the hosted on-link prefix. The MN, on receiving these RA messages, attempts to configure its interface by using either stateful or stateless address configuration modes. At the end of a successful address configuration procedure, the MN has one or more addresses derived from its home network prefix.

The LMA receives all packets sent to the MN from any nodes inside or outside the PMIPv6 domain. It forwards the received packets to the MAG through a bidirectional tunnel. After receiving the packets, the MAG removes the outer header and forwards the packets to the MN on its access link.

(2) Handover Control in PMIPv6

Figure 6.12 shows the handover signaling flow in PMIPv6. When the MN moves away from the previously MAG (p-MAG) and attaches to a new MAG (n-MAG), the p-MAG detects the leaving MN and sends a PBU message to the LMA to update the binding state. The LMA receives the PBU and returns a PBA message. After detecting the MN’s attachment, the n-MAG updates the binding and routing information of the LMA. The MN on the new link would receive the same RA message, believing that it remains on the same link and uses the same address.

After obtaining its initial address configuration in the PMIPv6 domain, if the MN changes its point of attachment, then the MAG on the previous link will detect the MN’s detachment from the link. It will signal the LMA and will remove the binding and routing state for that MN. The LMA, upon receiving this request, will identify the corresponding mobility session for which the request was received. The MAG on the new access link will then signal the LMA to update the binding state. After receiving the PBA message, a bidirectional tunnel is established.

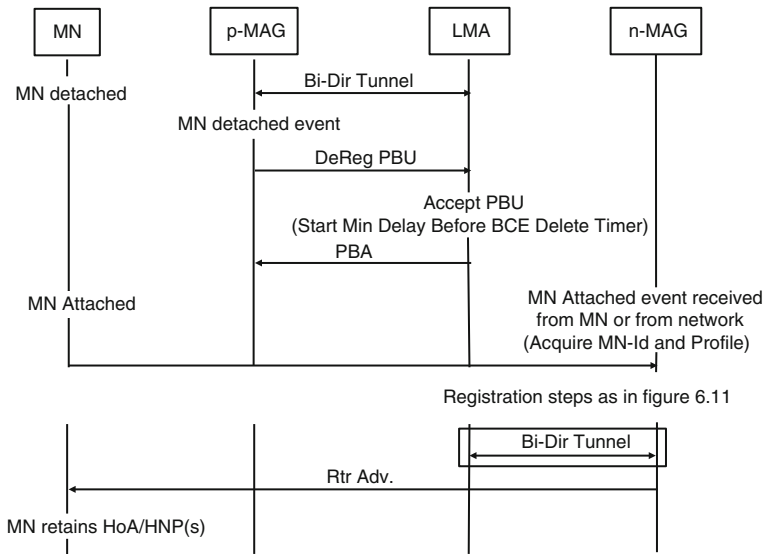


Fig. 6.12 Mobile node handover signaling flow in PMIPv6

After the bidirectional tunnel is established, the MN sends a router request message to the n-MAG. Then, the MAG sends a RA message containing the MN's home network prefix. This process establishes the connection between the MN and the MAG, and then, handover is accomplished.

6.2.3 Network Mobility

Since more hosts/devices became mobile, some of them may move together as a group, i.e., a mobile subnetwork, for example, lots of mobile phones and tablet pc within a bus or aircraft. This kind of subnet needs to connect to Internet to satisfy user requirements of those devices connected to it. Due to mobility, routers connect such subnet to the Internet need to change to new network attachment points.

The mobility of such a subnet complicates mobility management; for example, some hosts in a mobile subnet may not support MIP, or may not detect that they have already changed to a new network attachment point, or there may be some subnetworks within this moving subnetwork (i.e., nested networks).

Obviously, host-based and network-based mobility management protocol designed for single host mobility is not capable of handling above scenarios. In the year 2000, Motorola's Paris laboratory proposed to extend MIPv6 by establishing a tunnel between the HA and the Mobile Network Gateway, and on behalf of all hosts of the subnetwork, Mobile Network Gateway creates and updates bindings of location with the HA [17]. In the year 2001, NTT DoCoMo proposed to extend

MIPv6 by dynamic exchange routing information between mobile router and HA when this mobile router moves to a network attachment point [18].

In November 2002, IETF set up NEMO [19] working group to study how to support network mobility based on MIPv6. In January 2005, RFC 3963 [10] was published, for supporting network mobility. Later, some enhancement to network mobility has been proposed, and such routing optimization, and multi-homing support, would also be discussed in this section.

(1) **Network architecture of NEMO**

The architecture defined by the RFC 3963 NEMO is shown in Fig. 6.13. Within this architecture, the mobile network consists of one or several subnetworks that need access to the Internet via a mobile router, and the mobile network move as a whole. Due to mobility, the mobile network may change to new network attachment point. The mobile router (MR) serves as a default gateway for those hosts within such mobile network for accessing the Internet and performs mobility management-related function on behalf of those hosts.

In the network architecture given in Fig. 6.13, the mobile network node (MNN) may be a local fixed node (LFN, without any mobility within the mobile network, which is not changing its network attachment point) or an MN with local mobility restricted to the mobile network (LMN), or a visiting mobile node (VMN) from outside the mobile network. Those nodes could be a host, or a router.

If a VMN is an MR, then the mobile network connected to this MR is a nested mobile network, e.g., MR2 and the mobile network 2 shown in Fig. 6.13. The mobile network providing Internet access for another mobile network is Parent-NEMO, and

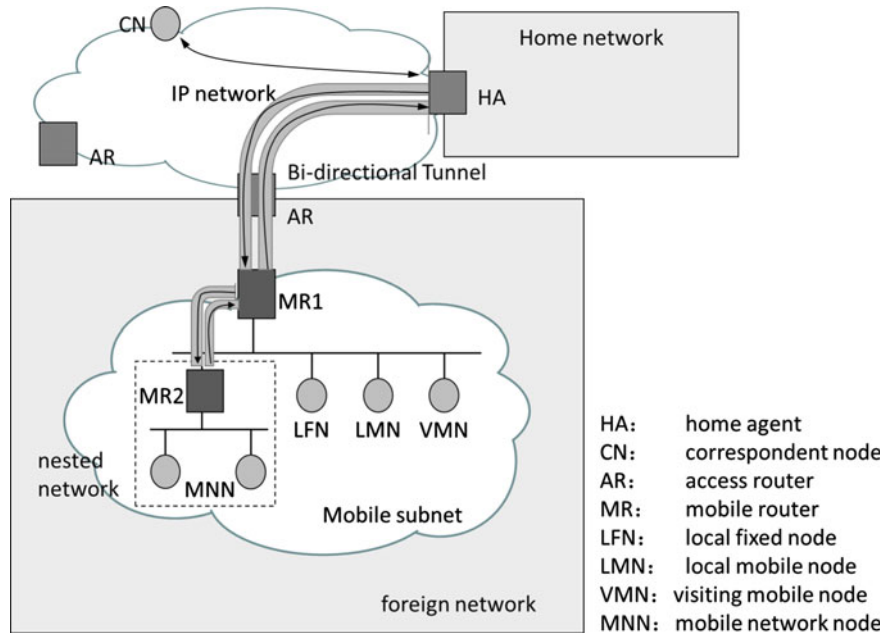


Fig. 6.13 Network architecture of IETF Network Mobility (NEMO)

the MR in the Parent-NEMO is Parent-MR, e.g., the MR1. The mobile network attached to another mobile network is Sub-NEMO and its MR is Sub-MR. The nested level of such a mobile network is not limited.

A bidirectional tunnel is established between the MR and the HA, to carry data packets to and from the HA and mobile network nodes. This is based on extending no routing optimization mode of MIPv6.

(2) **Basic Registration and data forwarding procedure of NEMO**

The MR can act as a mobile host or a mobile router. When acting as a mobile host, it is a host supporting MIPv6. When acting as a mobile router, it works as a border router, performs mobility management functions for hosts within its network, establishes and maintains the bidirectional tunnel with itself and HA, and encapsulates/de-encapsulates packets to and from HA and network, etc. In such case, the MR has at least two interfaces, one accesses the outside network, and the other is a default gateway for hosts within the mobile network.

The procedure of MR registration is shown in Fig. 6.14. When the mobile network leaves its home network and moves to another network attachment point, via an access router (AR) in a foreign network, the MR obtains the care-of address (CoA) from this AR and then sends a BU message to the HA to update the binding information. In explicit mode, one or more home network prefixes are included in the BU message, while in the implicit mode, no prefix is included, and HA needs to figure it out. This procedure is also applied to the nested mobile network. The Parent-MR acts as the AR which allocates CoA to the Sub-MR in the nested mobile network.

Within the BU message, a new flag, Mobile Router Flag [10] is set to indicate this BU is from an MR. A new option, mobile network prefix (MNP) option [10], is included to carry the home network prefix of the mobile network. If MR has more than one home network prefixes, then multiple AR options [10] are used. Upon the receipt of the BU, the HA updates its binding accordingly; that is, it stores the mapping between ARes and related CoA in the binding cache. And a new flag, Mobile Router Flag, is set to indicate supporting of mobile routers, in the replying BA message.

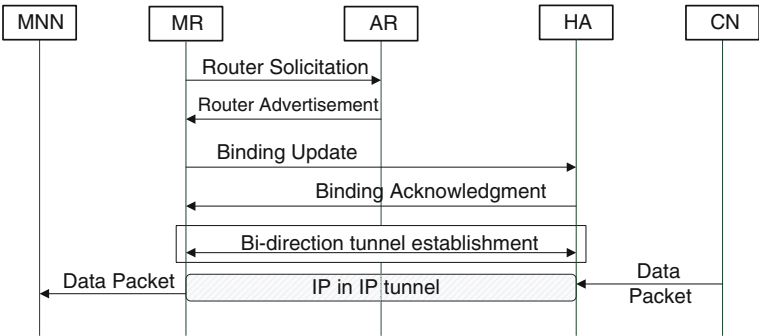


Fig. 6.14 Registration and data forwarding in Network Mobility (NEMO)

After each BU procedure, a bidirectional tunnel is established between the HA and the MR. When receiving packets from CN, HA forwards those packets via this tunnel to MR, if such packets are designated to home network prefixes of such mobile network. The MR encapsulates the packet and forwards via the interface where the mobile network is connected. On the other direction, the MR forwards packets from the mobile network to HA via such tunnel, and HA encapsulates packets and forwards them to CN.

For the nested mobile network case, let us take the example shown in Fig. 6.13, which is mobile network 2 with MR2. In this case, a bidirectional tunnel between MR2 and MR1 (as HA for MR2) is established after BU procedure between MR2 and MR1 is finished. For communication between MN within mobile network 2 and mobile network 1, data packets need to travel via the tunnel between MR2 and MR1. For communication between MN within mobile network 2 and CN located outside such foreign network, data packets need to pass tunnel between MR2 and MR1, and tunnel between MR1 and HA.

(3) Routing optimization of NEMO

Basic NEMO architecture uses bidirectional tunnel between HA and MR to transfer packet between CN and MN located in mobile network, which is the same as that how a packet is transferred in MIPv6 with no routing optimization. Since the de/encapsulation operation puts more processing burdens on HA, and increases transmission delay (please think about triangle problem), then HA becomes the bottleneck of data transferring if more mobile networks are registered. These problems are exacerbated for the nested mobile network case since multiple layer de/encapsulation operation is needed.

The basic idea of routing optimization is quite similar to that of MIPv6. The MR sends the BU message carries the Mobile Network Prefix Option directly to CN or router where CN connected, rather than HA. In this case, a bidirectional tunnel is established between MR and CN or router. Thus, HA is relieved from data forwarding task.

Another problem is a stalemate situation that appears when the Parent-MR visits one of its Sub-MR due to mobility. An example is shown in Fig. 6.15. The topology of mobile networks before the movement and the topology after movement is depicted. After movement, the previous Parent-MR MR1 connects to mobile network 2, which was its previous subnetwork. It is difficult to maintain tunnels for nested topologies in mobility, and BU/BA messages may get lost often. Thus, the NEMO-OLSR [20] proposed to use the modified mobile ad hoc network routing protocol OLSR [21], to make all MRs under the same AR networked as a mobile ad hoc network. Therefore, problem caused by nested mobile network does not exist anymore.

(4) Multi-homing in NEMO

If an MR has multiple egress interfaces to different network attachment points, or a mobile network has multiple MRs, or a mobile network is associated with

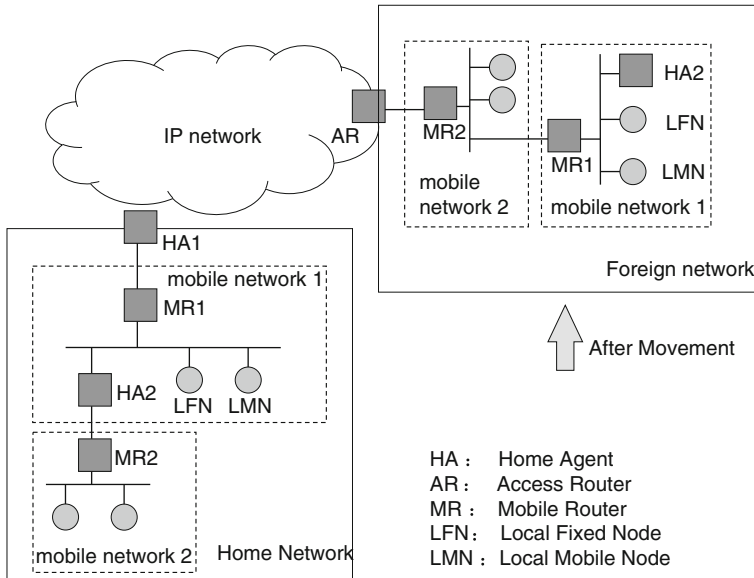


Fig. 6.15 The stalemate situation: an example

multiple HAs, or a mobile network has multiple global routable address prefixes, such a mobile network is considered as multi-homed.

In multi-homed mobile network, multiple data path exists. Both HA and MR need to select path for data transfer and detect path failure by lifetime [23] value contained in BU and BA is also needed. Based on its gathered information, such as cost and bandwidth of available paths and the preference of bidirectional tunnels between the MR and the HA obtained in the RAs, a new path may be selected by MR. For HA, it may select a path based on the information of the available MRs stored in the binding cache. The MR has multiple CoAs which binds to the same MNP by sending multiple binding identifier mobility options to the HA. Those bindings provide valuable information for HA to select data path.

Multiple HAs serving the same MNP may be located on different links. In order to allow only one HA to announce the MNP, Inter Home Agent Protocol (IHAP) [22] is used as a synchronization mechanism among HAs. Every HA is configured with a list of other HAs serving the same MNP. The Dynamic Home Agent Address Discovery mechanism [23] is used to choose one primary HA. The primary HA notifies the binding information to other HAs in the list by sending Binding Information Update [22] messages.

Multiple MRs broadcasting the same MNP may coexist at the same time. It is impossible for these MRs to bind the same MNP to the HA, because HA does not know through which MR the packets to be sent with destination address configured from the MNP. Token-based DND [24] is a possible solution to solve the problem. MRs supporting the same MNP are classified into owner and borrower. There is

only one owner, and it sends RA including the MNP and a token representing the MNP. If received the RA from the owner, an MR becomes the borrower of the MNP and sends the BU. Each BU includes a token and an indication to tell whether the MR is the owner or the borrower. When the HA sends packets with destination address configured from the MNP, it can choose either the owner or one borrower based on locally configured policies.

6.2.4 IP Flow Mobility

By using the MIP mechanism, the MN can maintain session continuity during mobility. The MIP only allows the MN to bind one single care-of address to its home address, which means that MN with multiple IP interfaces cannot utilize the two features: MIP and multiple interfaces, simultaneously. As a technology trend, the number of mobile nodes (MNs) with multiple interfaces is growing enormously. To extend the use of MIP, IETF developed RFC 5648 [25], which enhanced the original MIPv6 protocol to enable an MN to bind multiple care-of address to the same home address. RFC 5648 defines a new binding identification mobility option. Each such mobility option carries a Binding Identification Number (BID), which identifies a binding between a care-of address and the home address.

- A MN can register multiple bindings at one time by sending a single BU message, with multiple BID mobility options included. When a HA receives a BU with BID mobility option, it checks whether the BID contained in the BID mobility option is already included in its binding cache. If not, the HA copies the BID and the related care-of address and home address to the binding cache. A MN can also replace some or all of the bindings at the HA with new bindings by bulk registration. It should be noted that bulk registration is only supported for home registration but not for correspondent registration.

By using the above mechanism, HA maintains a binding cache which contains one or more entries, and each entry is identified by a BID and corresponds to a binding. The conceptual binding cache is illustrated in Table 6.2.

Based on RFC 5648, RFC 6089 [26] further extends MIPv6 and DSMIPv6 to allow special treatment of a particular IP flow, that is, to allow MN to bind a specific IP flow to a certain care-of address, which is called flow binding. Flow binding thus enables the HA to direct an IP flow to certain care-of address. Here, an IP flow is defined as a series of IP packets matching a traffic selector. Traffic selector, as defined in RFC 6088 [27], identifies the source and destination IP

Table 6.2 Binding cache maintained in home agent

Home address	Care-of address	Binding ID
HoA1	CoA1	BID1
HoA1	CoA2	BID2
...	...	...

Table 6.3 Flow binding entries maintained in HA

FID-PRI	FID	Traffic selector	BIDs	A/I
10	4	TCP	2	Active
30	2	srcAddr=IPy	4	Inactive
40	5	UDP	1,3	Active

addresses, source and destination port numbers, transport protocol number, and other fields in IP and higher-layer headers.

RFC 6089 introduced flow identification mobility option in MIPv6 and DSMIPv6. This new mobility option contains information that allows the receiver of a BU install policies on an IP flow and routes it to a certain care-of address. The information contained in flow identification mobility option includes a FID (flow identifier), traffic selector suboption, and binding reference suboption. The binding reference suboption field carries a BID to which a certain flow is corresponded, since a BID uniquely identifies a binding; thus, it can indicate the care-of address to which the IP flow is destined. When receiving a BU message containing flow identification mobility option, the HA stores the above information in its flow binding entries list, and each entry in this list includes the following parameters:

- FID: The FID uniquely identifies a flow binding entry for an MN.
- Traffic Selector: The traffic selector identifies a certain IP flow;
- BID: The BID to which the flow binding entry is associated.
- Active/inactive flag: This flag indicates whether the entry is active or inactive.
- FID-PRI: This field indicates the priority of the entry among all the entries associating with the same BID.

The following conceptual flow binding entries list is shown as an example in Table 6.3.

It can be seen that by using flow identification mobility option, routing of IP flow to different care-of address can be realized. The MN can at any time modify the flow binding by sending BU message to the HA, or to the correspondent node.

6.3 Key Control Technologies for Mobility Management at Network Layer

6.3.1 Overview

As illustrated in the mobility management reference model, the mobility management solutions at network layer realize mobility management functions at user data transport plane, control plane, and management plane. From the user data transport plane perspective, such a mobility management solution shields network-layer and lower-layer operation to higher protocol layers, such as application layer. From the control plane perspective, the mobility management solution at network layer has four functions, which are given in Table 2.8.

The MIP (MIPv4, MIPv6, and DSMIP) and PMIP are typical mobility management protocols at network layer. However, as the basis of mobility management protocols, there are some severe problems for MIP and PMIP, such as high signaling costs for location management, long delay, and packet loss for handover. Those would result in performance degradation and even communication interruptions; thus, optimizations are necessary. In this section, problems and solutions are discussed in terms of location management and handover control. Security aspect could be found later in this chapter.

6.3.2 Location Management and Its Optimizations

(1) Problems Statement

For the MIP, functions of location management are much similar to the cellular mobile communication system, i.e., network element knows the MN's location via registration or BU procedure. That is, when an MN changes its points of attachment, it will register its new CoA to its HA and/or to its CN if routing optimization is used. The MN also needs to periodically notify its current CoA to its HA and/or CN in order to keep its location information available. For PMIP, the network element can deal with MN mobility without the MN involvement. However, there is no paging mechanism in basic MIP and PMIP. The network element only relies on MN's location update to locate the MN.

(2) Basic ideas of Proposed Optimization Solutions

The basic idea is to learn location management mechanism from the cellular mobile communication system:

- (a) a region is used as the approximate location area of the MN, and such region is registered in a network element which performs location management functions; location re-registration procedure is needed if an MN moves out of its registered region;
- (b) the network element introduces paging mechanism at network layer, i.e., IP paging within the paging region where the MN is registered;
- (c) each network element manages certain amount of local location areas, and those network elements may consist of a hierarchical location management architecture, that is, if mobile node moves into a new region that belongs to the same network node as the old region belongs, the re-registration signaling will end in this network node; if the new region belongs to a different network node other than the old region, the re-registration signaling will go to a upper hierarchical network node.

In cellular system, location updating and reachable management mechanism work as an integrity location management solution, which consists of idea (a), (b), and (c) as a whole. However, there is an interesting difference from cellular system

that optimization solution for network-layer mobility appears in such an appearance by using one idea or combinations of above ideas, such as P-MIP, which just contains paging extension to MIP. This gives a more flexible way to deploy solutions. Related and protocol solutions detailed in this section could be refereed by readers with such “flexibility,” as well. However, due to complicated IP network, no perfect solution has been found to fit all possible scenarios. The main goal of every solution is to balance the amount of signaling used for location updating and registration between the costs for IP paging, at whole system level.

(3) Solutions mainly target to optimize Location Updating

Based on the scope of mobility, mobility management solutions can be classified into macro-mobility management, and micro-mobility management [28, 29]. The MIP and PMIP are typical macro-mobility management solutions that support global mobility of MN. On the contrast, the micro-mobility management solutions are optimized to handle micro-mobility within a single small region by limiting registration-related signaling within this area. For micro-mobility scenario, those so-called regional mobility management solutions not only reduce location management signaling load, but also decrease the handover latency and packet lose rate.

Typical micro-mobility management solutions include MIPv4 Region Registration (MIP-RR), Hierarchical Mobile IPv6 (HMIPv6), Cellular IP (CIP) [30], Telecommunications-Enhanced MIP Architecture for fast Intra-domain Mobility (TeleMIP) [31], and Handover Aware Wireless Access Internet Infrastructure (HAWAII) [32].

Depending on how user data packets are treated, mobility management solutions can be classified into routing-based and tunnel-based. The routing-based solution uses existing routing and forwarding functions to transfer packets for the MN, e.g., CIP and HAWAII, while the tunnel-based solution uses IP-in-IP tunnel to forward packets, such as MIP-RR [33] and HMIPv6 [34].

A. MIP-RR and Hierarchical Mobile IP

In MIP-RR [33], a network functional entity named as Gateway Foreign Agent (GFA) has been introduced into MIPv4, for micro-mobility management purpose. Some FAs are connected to the GFA directly within the certain location area managed by the GFA. The MNs register with HA by using GFA’s IP address as their care-of addresses. Data forwarding tunnels are established between GFA and its connected FAs. No location updates signaling to HA if MN does not move out of current location area managed by this GFA.

In HMIPv6 [34], a mobility anchor point (MAP) function entity has been introduced to enhance MIPv6. In a foreign network, the MN obtains a regional care-of address (RCoA) from MAP and an on-link CoA from a foreign subnetwork connected to this MAP. Only for MN mobility between different MAPs, Binding Update message containing RCoA address is sent to HA and corresponding node is needed. For MN moving from one subnetwork into a different subnetwork connected to a same MAP, Binding Updating procedure is performed to update the

binding cache in MAP with MNs’ new LCoA obtained in the new subnetwork. In this case, the MAP works as a user plane anchor for traffic to and from MN and corresponding nodes.

From the perspective of the HA and CN, hierarchical location management hides the MN’s micro-mobility within one regional location management area; thus, registration signaling related to such mobility is avoided, and handover latency for such mobility is also reduced.

B. TeleMIP

TeleMIP [31] is an extension of MIP inspired by mobility management schema of cellular mobile communication networks. In [TeleMIP], one or more FAs are connected to a newly introduced mobility management entity, TeleMIP Mobile Agent (TMA). For the first time registration of an MN, the TMA allocates a global care-of address for it, and this GCoA is registered in this MN’s HA. For mobility between base stations connected to the same FA, no registration procedure is needed and the local care-of address (LCoA) assigned by this FA is kept. For mobility that causes FA change, new LCoA will be assigned to the MN during the registration procedure running between the TMA, FAs, and the MN. The mobility scenarios without TMA change are transparent to the HA. The TMA is aware of the exact location of the MN and can route packets to this MN. Since LCoA is only used within the network domain managed by TMA, private IP addressing and routing algorithms could be used in this domain. Please refer to Fig. 6.16 for the TeleMIP network with three mobility management hierarchy.

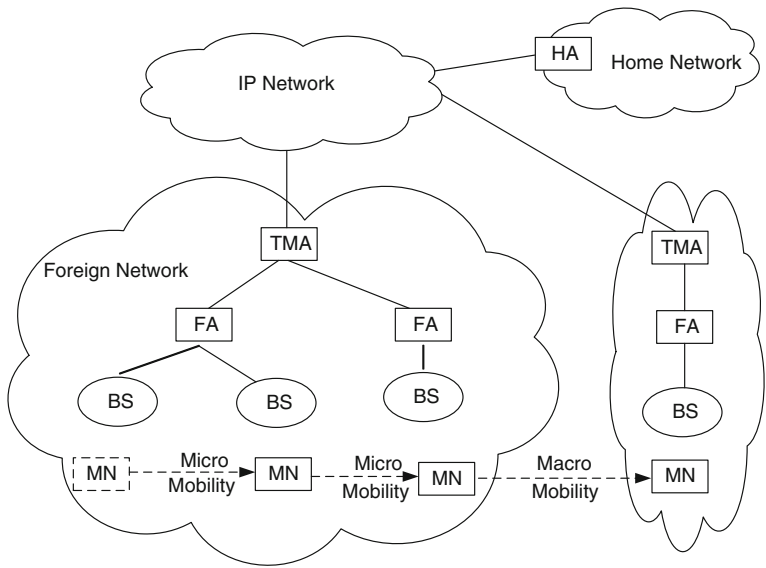


Fig. 6.16 The TeleMIP network architecture with three mobility management hierarchical

C. Cellular IP

The CIP [30] uses a different method to track MN location, i.e., regular data packets are sent to update host location information. That is, network nodes in a cellular IP access network monitor MN originated packets and maintain a distributed location database that is used to route packets to MNs. Thus, location update signaling is eliminated.

However, for mobility between different cellular IP access networks, a macro-mobility management protocol is needed to support global mobility, such as MIP.

D. HAWAII

The Handoff-Aware Wireless Access Internet Infrastructure (HAWAII) [32] relies on MIP for interdomain mobility management (i.e., macro-mobility management), and a separate routing protocol for intradomain mobility management (i.e., micro-mobility) was proposed. A MN entering a new FA domain is assigned a collocated care-of address, and this address is kept unchanged by the MN for mobility within this domain belonging to this FA. Within this domain, i.e., a HAWAII network domain, network nodes execute a generic IP routing protocol and maintain mobility-specific routing information as per host routes added to legacy routing tables. Thus, the existing packet forwarding function is reused for this MN. The mobile-specific routing entries for this MN are created, updated, and modified by explicit signaling messages. Four alternative path setup schemes are also defined for control handover between access points, to eliminate packet loss, minimize handover latency, and maintain packet ordering. Hawaii also uses IP multi-casting to page idle mobile hosts when incoming data packets arrive if there is no recent routing information found for this MN.

(4) IP Paging Solutions

RFC 3132 [35] describes the research scope for IP paging. RFC 3154 [36] describes the functional architecture for IP paging and requirements for IDLE mode MN. By introducing the IDLE mode, the MN saves lots of power since it can turn off the radio transmitter if there is no packet to transmit. Those two RFCs give guidance of research work for IP paging.

To lower the paging signaling cost and to shorten paging response time are two important goals that a network-layer paging solution pursues. Basically, the paging area construction, paging architectures (i.e., paging protocols and procedure initiator), and paging algorithms are three major aspects that each solution must consider and directly affect the efficiency and performance of solutions.

A. Paging extension for MIPv4

In a solution proposed in paging extension for MIPv6 [37], concreate paging area construction methods, procedures, and messages for IP paging have been proposed, by extending the MIPv4 protocol.

As proposed in this solution, if the MN has no data to transfer for certain time, it becomes an IDLE MN, i.e., it will not initiate normal registration procedure to this HA if it does not move out of current paging area. The paging area consists of several IP subnets that belong to the corresponding FAs. The correlation of a paging area and several FAs is pre-configured in each of the FAs. The IDLE MN detects whether it moves out of the current paging area by comparing the IP prefixes received from Agent Advertisement message with its care-of address, and if different, it assumes it has moved to another network. The lifetime value received in the message can also be used as an alternative: If there is no new Advertisement Message received before the expiration of the lifetime, it also assumes it has moved to another network, and then, MIP registration is initiated. For ACTIVE MN, it operates in exactly the same manner as in MIP. When a FA received IP data packets targeted to a care-of address, it will first judge whether the targeted MN is in active mode or not. If yes, the IP packets are decapsulated first and then forwarded to this MN. If not, the paging procedure is initiated.

A typical network structure of this solution is shown in Fig. 6.17; the corresponding paging procedure is shown in Fig. 6.18. In this example, paging area-1 contains three FAs, and an MN firstly registered via FA1 and then goes into an IDLE mode and moves within the paging area-1. Suppose that when data packets whose destination is this MN arrive at FA1 and this IDLE MN is within FA3's coverage. The FA1 first checks the status of this MN and turns out that this MN is not in active mode. Then, paging is initiated: FA1 knows that FA2 and FA3 are within the same paging area-1, so it sends paging request messages to FA2 and FA3 and sends a broadcasting paging message within its coverage; FA2 and FA3 also

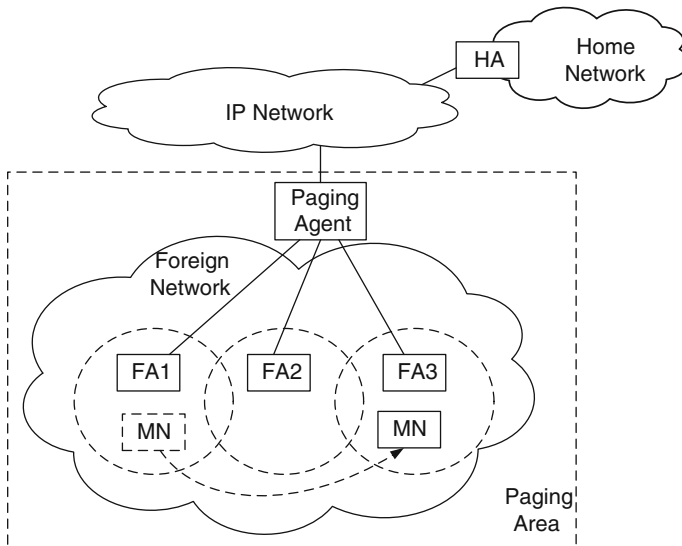


Fig. 6.17 An example network structure of paging extension to MIPMIP

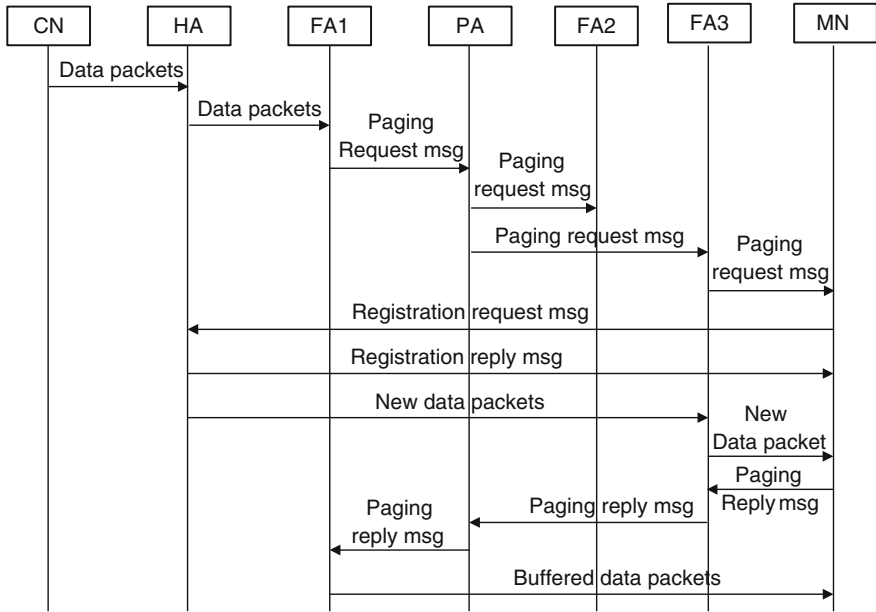


Fig. 6.18 An example paging procedure of P-MIP

send a broadcasting paging message within their coverage. In this case, FA1 performs the paging agent function defined in this solution. When this MN received paging messages, it registers with its HA via current visiting foreign agent FA3 and sends paging reply message to FA3. The HA replies the registration request and sends newly arrived data packets to the MN via FA3. The FA3 finds out that itself is not the previously registered FA of this MN and then forwards the paging reply message to the previously registered foreign agent FA1. Upon the reception of the paging reply, buffered data packets at FA1 are forwarded to this MN via FA3.

B. Dynamic Multi-step Paging scheme for PMIPv6

In the solution proposed in dynamic multi-step paging scheme for PMIPv6 [38], a paging area is constructed by an MAG based on MN's velocity and neighbor MAGs within n hop distance of this MAG. The Router Solicitation message contains velocity from the MN. The n is network configuration parameter. This paging area, which contains the address of several MAGs, is sent to LMA by extending PBU message and stored by LMA. Upon the reception of PBA message from LMA, the MAG sets *Hop-To-Paging area (HTP)* to n and included it in RA message sent to the MN. During this procedure, the network also acknowledges the MN that it is allowed to enter idle state. The HTP is included afterward in RA/RS message exchange performed when MN moves to a new MAG. If it is not zero, the PBU procedure will not be initiated by the new MAG, HTP is decreased by one and sent back to the MN in RA message. Once it reaches to zero, a new paging area will

be reconstructed by the MAG participated in next round of RA/RS message and is sent to LMA in the PBU message. Such a design gives the possibility to save signaling related to proxy binding procedures.

When downlink data arrive at LMA, the LMA performs H multi-step paging, that is, paging the MN in several incremental steps instead of flooding paging messages across the entire paging area stored in LMA. In the first step paging, the actual paging area only consists of some MAGs from the stored paging area. If no reply is received, the LMA increases the paging area to contain more MAGs and pages in such increased area. The paging area used in each step is calculated by LMA by using H as the input. The numerical and simulation results showed that in the PMIPv6 networks, this paging scheme can reduce the signaling cost for IP paging significantly and satisfy the maximum tolerable paging delay requested by the network.

6.3.3 *Handover Control and Its Optimization*

(1) **Problems Statement**

For MIP, the handover procedure consists of the following steps from network-layer perspective: (a) movement detection, that is, the MN finds out it has moved into a foreign network, (b) the MN gets a new CoA at the foreign network and registers its new CoA to the HA (and/or CN), and (c) upon the reception of the new CoA, the HA updates its binding entry and redirects the packet forwarding. Due to the movement detection relying on IP layer message exchange, the CoA configuration needs duplicated address detection check, and registration signaling may across long distance to complete the handover procedures. Once the mobile moves out from this old subnetwork, packets that are sent to its old CoA address are not deliverable and will be discarded by old subnetwork, until handover complete. For Proxy MIP, the movement detection via network-layer mechanism also contributed lots to the handover delay. However, time used for registration signaling becomes very small since that signaling is localized between MAG and LMA.

The advantage of network-layer handover is that such mechanisms are agnostic to radio-layer technologies and thus easy to fulfill vertical handover. But due to lack of lower layer's help, the good performance of handover such as low latency and low packet loss rate cannot be achieved. The long handover latency and packet loss are unacceptable to real-time services and packet loss-sensitive applications, respectively.

(2) **Basic ideas of Proposed Optimization Solutions**

Lots of solutions are proposed to minimize the handover delay and packet loss. The basic ideas of those solutions include (a) use of cross-layer design method, which combines link-layer and network-layer mechanism to accelerate movement detection, (b) reduce registration signaling by completing such registration at the old

subnetwork or in advance to move to a new subnetwork, for example, the new CoA address configuration at the old subnetwork, pre-registration/BU, etc., and (c) cache and forward packets received by old subnetwork to new subnetwork, to reduce packet loss rate, and in the meantime, packet delivery ordering function regarding both forwarded packets and the new packet arrived at new subnetwork is also needed.

From the performance perspective, handover could be classified into fast handover, which means handover latency is minimized; smooth handover, which means no packet loss during the handover procedure; and seamless handover, which means both handover latency and packet loss rate are minimized to the extent that such handover is transparent to upper-layer application.

(3) Typical Handover Solutions for Mobile IP

A. Low latency handover for Mobile IPv4

In the low latency handover for MIPv6 [39], two handover methods, pre-registration and post-registration, are proposed. The pre-registration allows MN to “pre-build” its registration state on the new FA prior to the occurrence of an underlying L2 handover. Both the MN and FA can trigger network-layer handover by using layer 2 trigger; thus, seamless handovers between FAs could be achieved. This method relies on MIPv6 security. The post-registration allows old and new FAs to utilize layer 2 that triggers to set up a bidirectional tunnel between them, which allows the MN to continue using its old FA once it accesses the new FA’s subnet. Thus, the impact on real-time applications is minimized. However, the MN must eventually perform a formal MIPv6 registration, even if this could be delayed by the MN or foreign agent. Moreover, the pre-registration and post-registration handover methods could be executed in parallel, which turned into a combined method.

B. Mobile IP Fast Handovers

In fast handovers for MIPv4 (RFC4998 [40]) and MIPv6 (RFC5568 [41]) basically using the same idea for handover optimization, MIPv6 is taken as an example in the following discussion.

Two fast handover procedures, namely predicative and reactive, are shown in Fig. 6.19.

In FMIPv6 [42], the cross-layer method is used to reduce movement detection latency. The MN is provided with [Access Point-ID, AR-Info] tuple information at its current connected AR (i.e., previous access router (PAR) in the above figure), upon requesting of such information with AP-IDs scanned by MNs’ link-layer function. That is, the MN discovers the available access points using link-layer functions and then requests subnet information corresponding to one or more of those discovered access points by performing router discovery at any time via the current connected router. The “Router Solicitation for Proxy Advertisement (RtSolPr)” and “Proxy Router Advertisement (PrRtAdv)” messages are used for such router discovery procedures. During this procedure, the MN also formulates a prospective new CoA (NCoA) when it is still connected to its PAR. When an MN is attached to a new access point which AP-ID is within its stored tuples, it knows the

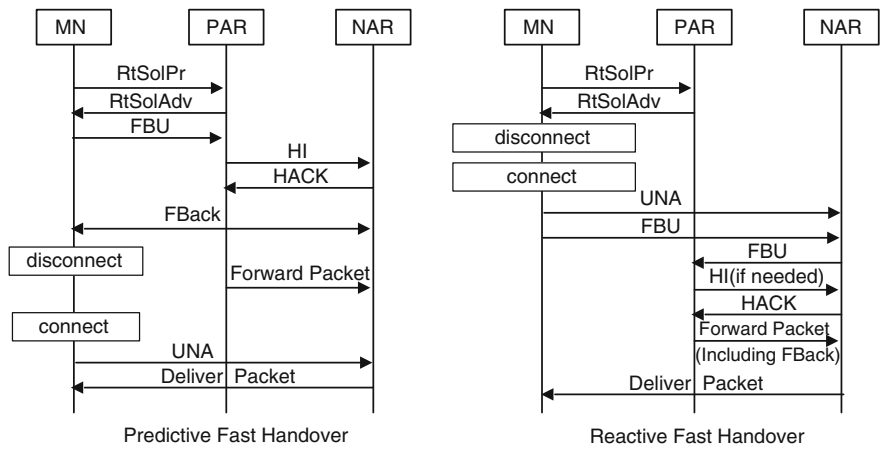


Fig. 6.19 Fast handover MIPv6 (FMIPv6)

corresponding new router’s subnetwork prefix, i.e., it has moved to a new subnetwork.

In order to use the NCoA immediately when it attaches to a corresponding subnetwork, an MN should send a “Fast Binding Update” at the current subnetwork prior to its moving to a new subnetwork and may receive the “Fast Binding Acknowledgment (FBack)” from PAR. This is the predicative fast handover. The FBack is also used by new access router (NAR) to indicate the acceptance of the NCoA. Even without receiving an FBack, the MN can still start using the NCoA after announcing its attachment through an Unsolicited Neighbor Advertisement (UNA) message (with the “O” bit set to zero, defined in RFC4861 [43]). Thus, the NCoA configuration latency is reduced.

The “FBU” sent to an MN’s PAR also triggers the establishment of a data forwarding tunnel between the previous CoA (PCoA) and NCoA. When feasible, the MN sends an FBU from the PAR’s link to establish the tunnel. Otherwise, the MN should send the FBU immediately after detecting attachment to the NAR, and this is reactive fast handover. In any case, this tunnel is used by PAR to tunnel packets arriving for PCoA to NCoA, while in the opposite direction, the MN sends packets with PCoA as source address to the PAR via this tunnel. Such a tunnel remains active until the BU procedure is completed. By using this tunnel, packet loss rate is reduced and packets forwarding is expedited.

For both kinds of fast handover procedure, the “Handover Initiate (HI)” message is used by AR to initiate the actual handover procedure at the network side, and “Handover Acknowledge (Hack)” corresponds to acknowledge messages.

(4) Typical Solutions for Proxy Mobile IP

A. Fast Handovers for Proxy Mobile IPv6

The fast handovers for PMIPv6 (RFC5949) [43] basically extend FMIPv6 [41] operation to minimize handover delay and packet loss, as well as to transfer

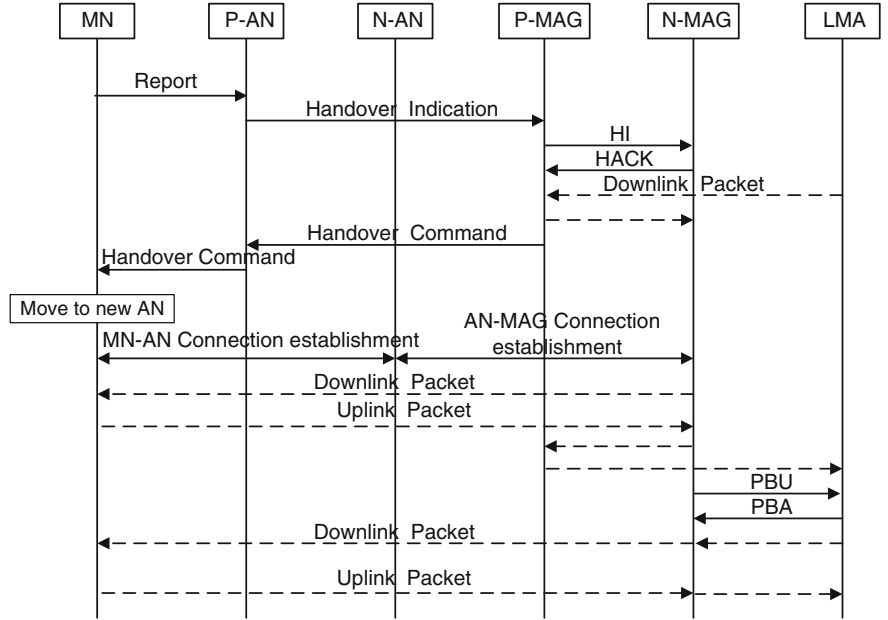


Fig. 6.20 Fast handover PMIPv6—predicative

network-resident context for an MN. Thus, there are two fast handover procedures, namely predicative and reactive, which are defined for PMIPv6. The predicative fast handover procedure, in which the initiator is the previous MAG, is shown in Fig. 6.20. In the reactive fast handover procedure, the initiator is the new MAG.

Unlike MIPv6, the MN in the PMIPv6 domain is not involved in network-layer mobility management procedure. However, in order to make the predictive fast handover to work effectively, MN involvement rather than purely relying on network-side mobility management operation is required. That is, the MN detects that a handover is imminent and reports its identifier (MN ID) and the new access point identifier to which the MN is most likely to move to the current attached AR (i.e., previous mobile access gateway (p-MAG)) via current access network (i.e., previous access network (p-AN)). The p-MAG derives the New Mobile Access Gateway (n-MAG) from this received new AP ID and then sends the HI message to this new MAG. The new MAG replies with HACK message back to p-MAG and requests p-MAG to buffer downlink packets. A bidirectional tunnel is established between p-MAG and the new MAG, to be used for forwarding downlink packets by p-MAG to n-MAG. The n-MAG will buffer received tunneled packets after encapsulation. Once the handover is ready on the network side, the MN is commanded to perform handover to the new access network by using access-specific link-layer technology. The connection between the MN and the new access network and the connection between the new access network and n-MAG are established after the MN attaches via the new access network to n-MAG. Then, n-MAG starts

to forward those buffered packets to MN. The n-MAG sends PBU to LMA. Before receiving the PBA, the uplink packets received by n-MAG are tunneled to p-MAG and sent to LMA. After that, downlink and uplink packets are transferred via n-MAG.

By reusing ideas from FMIPv6, the handover latency and packet loss are reduced. But the success rate of the handover procedure depends on several factors. Above is the predictive HO procedure in which the procedure starts from MN reporting link-layer information to p-AN, and the initiation of the actual handover is decided by a lower-layer method. For example, by measuring the radio signal strength value (RSS) and if such a RSS value meets the predefined threshold, the HI is sent to p-MAG to trigger subsequent handover procedure. However, the HI may be false due to the incorrect calculated RSS values, e.g., based on false information due to complicated radio environment, such as in the presence of obstacles and unpredictable mobility pattern of the MN.

B. Other Solutions

In packet lossless PMIPv6 (PL-PMIPv6) [44], in order to reduce handover latency and packet loss, the n-MAG registration with LMA is performed in advance of the MN movement by the p-MAG on behalf of the n-MAG, and buffering mechanism is utilized by the n-MAG, correspondingly. The success of this method depends on correctness of the prediction of the neighboring MAG that the MN is moving to.

In PMIPv6 with bicasting (PMIP-BI) [45], a new tunnel between n-MAG and LMA is established at the very early stage to prevent packet lost. That is, upon receiving the Link-Up trigger message, the p-MAG sends the HI message that contains the MN's proxy-CoA and HoA (MN-HoA), to the n-MAG. The n-MAG establishes the tunnel with LMA by sending the MN-HoA and MN identifier encapsulated in a PBU message. Upon receiving the PBU message, the LMA updates the BCE and transmits data packets to the MN via both p-MAG and n-MAG, simultaneously. Once the MN connects to n-MAG, the MN will receive those packets. Upon requested by n-MAG, the LMA stops packet transmission to the p-MAG eventually. This method is further enhanced in PMIPv6 with partial bicasting [PMIP-BIP] [46], where the n-MAG will buffer packets from LMA until it knows that the MN is connected via n-MAG. Then, buffered packets are transmitted to the MN by n-MAG, and n-MAG will request p-MAG to stop bicasting data packets.

6.3.4 Security Mechanism

Security threats faced by network-layer mobility management protocol come from the following aspects: man-in-the-middle attack, denial of service, impersonation attack, and attacks targeted at open wireless radio link. For example, signaling

messages such as agent discovery message of MIPv4 could be forged by malicious mode, which would make the MN misunderstood current binding is wrong; the man-in-the-middle attacker may send a fabricated BU message to the HA or the CN in order to pretend to be the MN; thus, packets can be redirected to this attacker, which would lead to the communication hijacked; impersonation attack and man-in-the-middle attack may also happen between the MAG and the LMA in PMIPv6 case. The attackers could trace the user's location simply by eavesdropping location update or BU messages, which would threaten the user's privacy. The data transferred on the wireless link may be eavesdropped or manipulated, by attackers. However, security protections for wireless link would not be discussed in this section since they are mainly related to specific radio access technologies.

In order to protect the signaling of mobility management protocols at the network layer as well as the user data and privacy, and prevent network resources from unauthorized usage, corresponding security mechanisms are proposed in the following way: Existing IP layer security mechanism such as IP Sec is enhanced to protect signaling and user data regarding the mobility and the existing AAA (authentication, authorization, and accounting) protocols are integrated with the mobility management protocol to provide authentication, authorization, and accounting functions for the MN and improve the system efficiency at the same time.

(1) IPSec based security mechanism

IPSec [47] is of the important part of the security mechanisms at network layer, which provides end-to-end security protection between a pair of communication parties.

The IPSec provides two basic security mechanisms: authentication and encryption. Authentication enables the receiver to verify or authenticate the true identity of the sender and thus knows that whether the received data have suffered malicious modification or not during transmission. Encryption is used to cipher user data, which can provide data confidentiality and integrity protection, prevent eavesdropping, and replay attack. The IP Sec protocol suite contains an authentication header (AH), an encapsulating security payload (ESP), and an Internet Key Exchange (IKE) protocol [48]. The AH is used to provide integrity and data origin authentication; the ESP defines methods for data encryption, and optional authentication is used to guarantee reliability. In practical IP communications, the selection of AH and ESP depends on the actual security requirements. Both AH and ESP can provide authentication protection, and selecting one of them for usage depends on the requirements. The IKE, as its name implied, is used to exchange keys, as well as establish and maintain security association (SA) between two parties. The SA is required for both AH and EPS. It is the fundamental part of IP Sec. One SA is a one-way logical connection; thus, security communication between the two parties requires a pair of SAs for sending and receiving directions. Two types of SA are defined: the transport mode SA and the tunnel mode SA. In the former, the AH or ESP headers appear after the IP headers and some selected

extension header, before any next layer protocols. In the latter, the AH or ESP headers appear after the outer IP header, which specifies the source and destination for the IP Sec processing, before the inner IP header indicating the actual source and destination for the packets.

In order to provide encryption and integrity protections for MIP signaling messages, IP Sec is used with some dedicate extensions. In MIPv4 [6], in order to authenticate the MIPv4 message is sent by the correct entity and is not modified during transmission, the MN, the FA, and the HA must support a Mobility Security Association (MSA) when an MSA is established between the MN and the HA; if the MN sends an MIPv4 message to the HA, it must include the SPI of the MSA and an authenticator calculated using the MSA. When the HA receives the message, it uses the SPI included in the message to find the context of the MSA and then calculates an authenticator using the found MSA and compares the calculated authenticator with the authenticator included in the received message. If they are the same, the HA considers the message is valid. Three extensions are defined to protect Registration Request and Registration Reply messages, i.e., Mobile-Home Authentication Extension, Mobile-Foreign Authentication Extension, and Foreign-Home Authentication Extension. The Mobile-Home Authentication Extension is used between the MN and the HA and must be included in the MIPv4 messages. The other extensions are optional.

MIPv6 [8] protocol defines a new next header, i.e., the mobility header, to transfer those messages used to create and manage bindings. The IPSec architecture is enhanced to support such mobility headers. The MH message type is defined as a selector to differentiate mobility header messages.

The PMIPv6 uses IPSec ESP transport mode between Mobile Access Gateway (MAG) and LMA to protect PMIPv6 signaling. IKEv2 [48] is used to set up SA between MAG and LMA.

(2) **MOBIKE**

When an IPSec SA is established between two nodes which support mobility or multi-homing, if the address of one of the nodes is changed or the addresses of both of the nodes are changed, and if the IP addresses used to establish the SA are changed because of the movement of one node, the IPSec SA needs to be re-established which will increase the transmission delay and may require user interaction. In order to solve these problems, MOBIKE [49] is designed to update the IP addresses associated with the IPSec SAs. In the MOBIKE, the party initiates the first IKE_SA is identified as initiator, and the other peer is responder. During IKE initial message exchanges, the support of MOBIKE and the available IP addresses are exchanged between the initiator and the responder, indicating to each other the support of MOBIKE and the available addresses if multi-homing is also supported. The initiator decides which address pair is used in the IPSec SA and initiates the change of the address pair to the responder by sending an INFORMATIONAL request to the responder. If the address pair is acceptable to the responder, it updates the addresses of the SA and returns an INFORMATIONAL

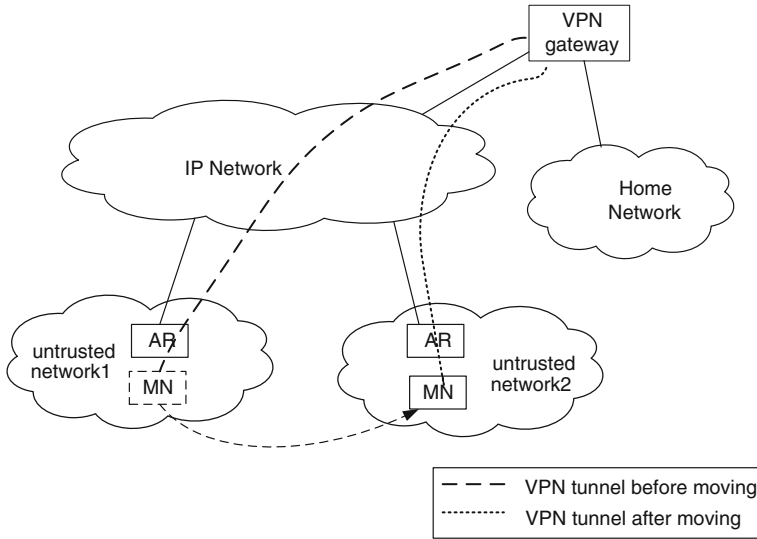


Fig. 6.21 An example of MOBIKE

response to the initiator. The responder can also initiate the INFORMATIONAL change to the initiator to update address pair of the IP Sec SA stored in the initiator.

An example for MOBIKE deployed in the enterprise networks and using MIPv6 and IPSec VPN to provide roaming and secure connectivity is shown in Fig. 6.21. When the MN roams to the untrusted network-1 outside the enterprise, the MN obtains an IP address from the untrusted network-1, which is used to set up a VPN tunnel to the VPN gateway. After the VPN tunnel is established, the VPN gateway allocates an IP address to the MN. The MN uses the address as the CoA and binds it to the HA. Then, the MN moves to the untrusted network-2, which allocates a new IP address to the MN. At this time, if the VPN gateway is not changed, the MN just has to use the MOBIKE protocol to update the VPN of its current address, i.e., the address allocated by the untrusted network-2. Thus, those SAs are moved smoothly rather than re-established. If the VPN gateway is changed, the MN has to establish a VPN tunnel with the new VPN gateway, obtaining a new care-of address from the VPN gateway and updating the HA.

(3) AAA-based security mechanism

AAA is an important mechanism for network security and also can be used to record the usage of the network resource. Currently, RADIUS (Remote Authentication Dial In User Service) [50] and Diameter [51] are two widely used AAA protocols.

RADIUS is a client/server protocol that can be used to verify the username and the password and deliver the service information to users. Although the RADIUS is simple, extensible, and widely used, it still has some drawbacks, such as using UDP

as the transport protocol, simple packet loss control mechanism, lack of retransmission mechanism, and centralized accounting service. Along with the network evolution, the AAA networks scale up, and so does its complexity, and the RADIUS cannot meet the new requirements anymore; thus, Diameter protocol is designed. Diameter protocol uses TCP and SCTP as the transport protocol to ensure the reliability of message transmission. It works as the foundation of AAA solutions to support MIP, network access control service, etc.

When the MN connects to an access network, before providing access to the MN, the access network will request the MN to provide credentials which is used by the AAA server in the home network to authenticate this MN. Once authenticated and authorized, the MN begins to access service via the access network. In order to support charging, the access network may start to collect the resource used by the MN.

In order to perform security-related function for MN using MIPv4 protocols, a general model based on AAA is described in RFC 2977 [52]. An example of the possible deployment is shown in Fig. 6.22. In this example, the FA and the HA connect to different AAA servers.

When the MN connects to the FA, it sends a Registration Request message to the FA. The FA initiates the authentication and authorization procedure to the AAA server in the visited domain by sending an AA-Mobile-Node-Request [53] message including the Registration Request [6] message. The message will be routed to the AAA server in the home domain, which authenticates the Registration Request. After successful authentication, if the information of a specific HA is not included in the Registration Request message, the AAA server allocates an HA to the MN.

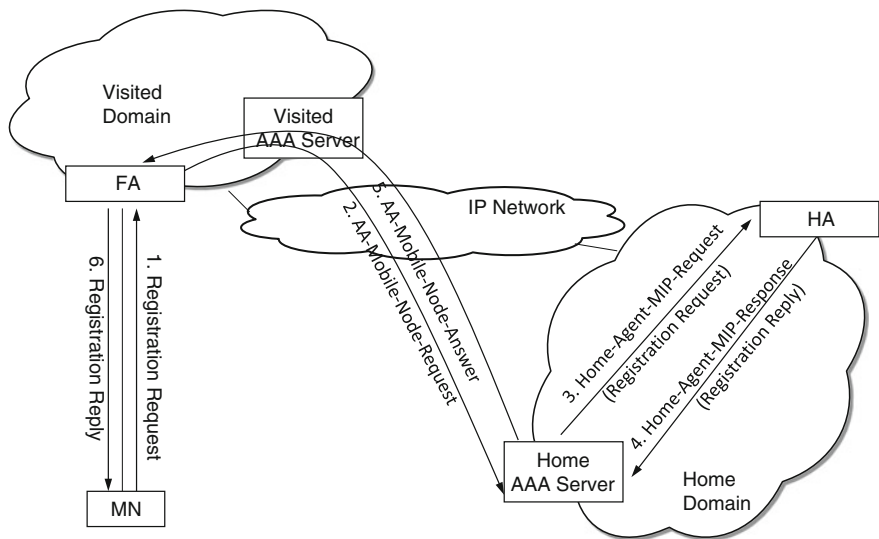


Fig. 6.22 Deployment example of interworking between AAA and MIPv4/MIPv4

Otherwise, the HA indicated by the UE is selected. Then, the AAA server sends the Home-Agent-MIP-Request [53] message containing the Registration Request message [6] to the HA. The HA processes the Registration Request and returns a Home-Agent-MIP-Response message including the Registration Reply [6] message. The AAA server in the home domain sends an AA-Mobile-Node-Answer [53] message to the FA via the AAA server in the visited domain, and the message includes the Registration Reply message along with the address of the HA. The FA sends the information to the MN.

In MIPv6, in order to register with the correct HA, the MN has to know information used for registration, such as the home link prefix and HA address. The information can be statically provisioned in the MN, but this not only increases the maintenance burden but also incurs security problems if the MN is lost. In order to solve such problems, i.e., how to provide information to MN in a secure manner is proposed by using interworking between MIPv6 and RADIUS [54], and between MIPv6 and Diameter [55]. Those two interworking mechanisms are similar, and the former is described here as an example.

There are two possible deployment scenarios. In the integrated scenario [54], MIPv6 bootstrapping is part of the network access authentication procedure, and the RADIUS server/client provides home registration information to the MN during this authentication procedure. When the MN connects to the access network, the network access server initiates the authentication procedure (e.g., EAP procedure), which is performed between the MN and the remote RADIUS server. In this example, the DHCP relay and the RADIUS client function are collocated with the network access server. If the network access server does not indicate the use of the local HA assignment, the remote RADIUS server selects an HA for the MN and sends such home registration information back in an Access-Accept [10] message to the network access server. If the use of local HA assignment is indicated, the remote RADIUS server will not provide the home registration information to the network access server. Then, the network access server is responsible to select a local HA for the MN. After the successful network access authentication, the MN sends a DHCPv6 request to the DHCP relay collocated with the network access server. When the network access server receives such a request, it forwards it together with the home registration information obtained during the authentication procedure, to the DHCPv6 server. DHCPv6 server sends DHCPv6 a reply including such home registration information to the MN. In the split scenario [47], the home registration information is provided by the HA to MN after the successful access authentication procedure.

During the IKEv2 procedure between MN and HA, the EAP procedure between the MN and the RADIUS server is performed. Such EAP messages are piggy-backed in the IKEv2 messages. By monitoring the EAP procedure, once HA knows that the EAP procedure is completed successfully, the HA sends the home registration information to the UE during the IKEv2 procedure.

Procedures for the above two scenarios are shown in Figs. 6.23 and 6.24, respectively.

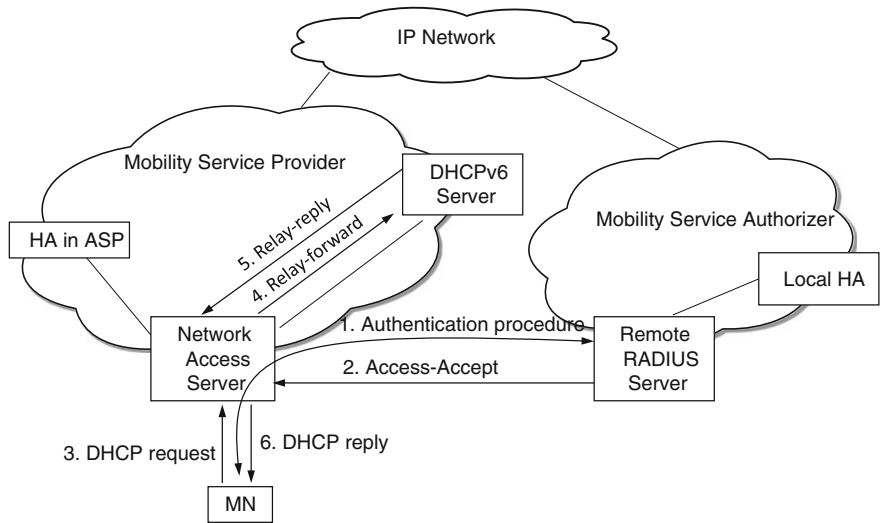


Fig. 6.23 Example of integrated scenario

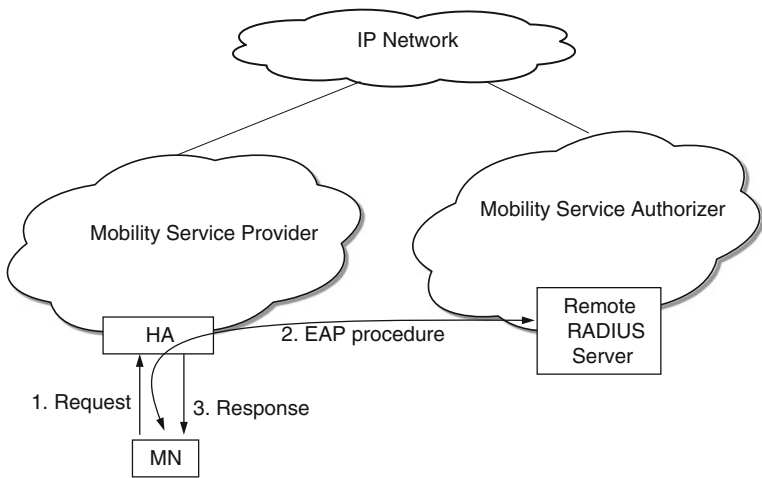


Fig. 6.24 Example of split scenario

In PMIPv6, RADIUS/Diameter protocol can be deployed between the MAG and the AAA server and between the LMA and the AAA server [55, 56]. When the MN attaches to an access network, the RADIUS client collocated with the MAG initiated the authentication procedure. After the successful authentication, the MAG receives the MN’s policy profile from AAA server. Then, the MAG sends a PBU message to the LMA. The LMA initiates the authorization procedure to active the mobility service. If the authorization succeeds, the LMA receives the MN’s policy profile.

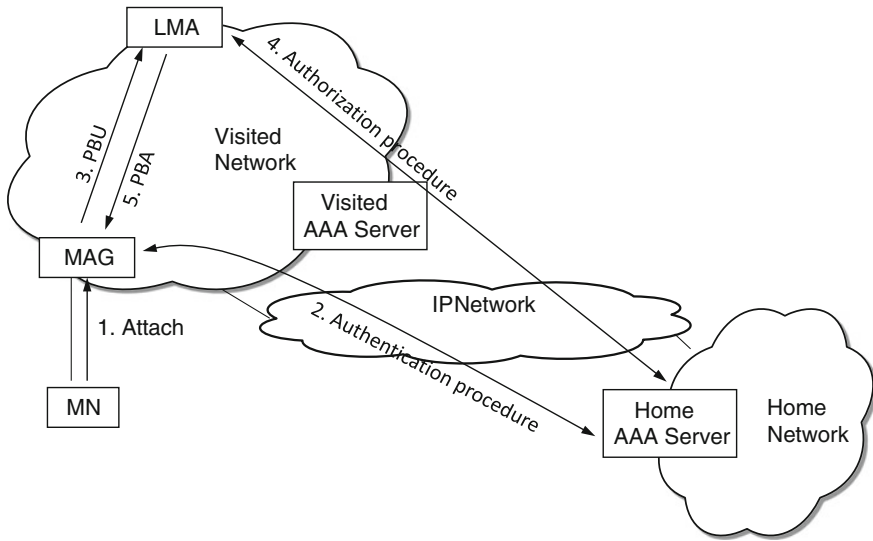


Fig. 6.25 An example of RADIUS support for PMIPv6

If LMA and MAG are located in a visited network or a local network, they may connect to the AAA server in the home network via the AAA server in the visited network or the local network. An example is shown in Fig. 6.25.

References

1. Postel J (1981) Internet protocol. IETF RFC 791
2. Droms R (1993) Dynamic host configuration Protocol. IETF RFC 1541
3. Postel J (1981) Transmission control protocol. IETF RFC 793
4. Perkins C (1996) IP mobility support for IPv4, IETF RFC 2002
5. Perkins C (2002) IP mobility support for IPv4, IETF RFC 3344
6. Perkins C (2010) IP mobility support for IPv4, Revised, IETF RFC 5944
7. Johnson D, Perkins C, Arkko J (2004) Mobility support in IPv6, IETF RFC 3775
8. Perkins C, Johnson D, Arkko J (2011) Mobility support in IPv6, IETF RFC 6275
9. Gundavelli S (ed), Leung K, Devarapalli V (2008) Proxy mobile IPv6. IETF RFC 5213
10. Devarapalli V, Wakikawa R, Petrescu A, Thubert P (2005) Network mobility (NEMO) basic support protocol. IETF RFC 3963
11. IETF RFC 5648 (2009) Multiple care-of addresses registration
12. Ke Pei, Jiandong Li, Haixia Dong, Feng Guo (2001) Study of the IP mobility. J Xidian Univ (Nat Sci) 28(5):663–667. doi:[10.3969/j.issn.1001-2400.2001.05.027](https://doi.org/10.3969/j.issn.1001-2400.2001.05.027)
13. Ioannidis J, Duchamp D, Maguire G (1991) IP-based protocols for mobile internetworking. Proc Conf Commun Archit Protoc ACM 21(4):235–245. doi:[10.1145/115994.116014](https://doi.org/10.1145/115994.116014)
14. Teraoka Fumio, Tokoro Mario (1993) Host migration transparency in IP networks: the VIP approach. ACM Sigcomm Comput Commun Rev 23(1):45–65. doi:[10.1145/173942.173946](https://doi.org/10.1145/173942.173946)
15. Rckhter U, Perkins CE (1992) Optimal routing for mobile host using IP's loose source route option. IETF Interent Draft

16. Soliman H (ed) (2009) Mobile IPv6 support for dual stack hosts and routers (DSMIPv6). IETF RFC 5555
17. Ernst T, Castelluccia C, Hong-Yon L (2000) Extending mobile IPv6 with multicast to support mobile networks in IPv6. In: IEEE 1st European conference on universal multiservice networks, pp 114–121. doi:[10.1109/ECUMN.2000.880732](https://doi.org/10.1109/ECUMN.2000.880732)
18. Okajima I, Umeda N, Yamao Y (2001) Architecture and mobile IPv6 extension supporting mobile networks in mobile communications. IEEE Proc Veh Technol Conf 4:2533–2537. doi:[10.1109/VTC.2001.957207](https://doi.org/10.1109/VTC.2001.957207)
19. <http://www.ietf.org/>. Accessed 23 Mar 2014
20. Ma X, Zhang J, Jia J, Liu Y (2011) The MANEMO route optimization in the rescue operation, draft-maxiaolei-manemo-ro-00.txt
21. Clausen T, Jacquet P (2003) Optimized link state routing protocol (OLSR). IETF RFC 3626
22. Wakikawa R, Devarapalli V, Thubert P (2004) Inter home agents protocol (HAHA), draft-wakikawa-mip6-nemo-haha-01.txt
23. Dupont F, Combes JM, Laurent-Maknavicius M (2008) Dynamic home agent address discovery (DHAAD) considered harmful, draft-dupont-mext-dhaadharmful-00.txt
24. Kumazawa M, Watanabe Y, Matsumoto T, Narayanan S (2005) Token based duplicate network detection for split mobile network (Token based DND), draft-kumazawa-nemo-tbdnd-02.txt
25. Wakikawa R (ed), Devarapalli V, Tsirtsis G, Ernst T, Nagami K (2009) Multiple care-of addresses registration. IETF RFC 5648:
26. Tsirtsis G, Soliman H, Montavont N, Giarretta G, Kuladinithi K (2011) Flow bindings in mobile IPv6 and network mobility (NEMO) Basic Support. IETF RFC 6089
27. IETF RFC 6088 (2011) Traffic selectors for flow bindings
28. Campbell AT, Gomez J, Sanghyo Kim et al (2002) Comparison of IP Micromobility Protocols. IEEE Wirel Commun 9(1):72–82. doi:[10.1109/MWC.2002.986462](https://doi.org/10.1109/MWC.2002.986462)
29. Saha D, Mukherjee A, Misra IS et al (2004) Mobility support in IP: a survey of related protocols. IEEE Netw 18(6):34–40. doi:[10.1109/MNET.2004.1355033](https://doi.org/10.1109/MNET.2004.1355033)
30. Campbell AT et al (1999) An overview of cellular IP. IEEE Wireless Commun Netw Conf (WCNC) 2:606–610. doi:[10.1109/WCNC.1999.796715](https://doi.org/10.1109/WCNC.1999.796715)
31. Das S et al (2000) TeleMIP: telecommunications-enhanced MIP architecture for fast intra-domain mobility. IEEE Pers Commun 7(4):50–58. doi:[10.1109/98.863996](https://doi.org/10.1109/98.863996)
32. Ramjee R et al (2002) HAWAII: a domain-based approach for supporting mobility in wide-area wireless networks. Netw IEEE/ACM Trans 10(3):396–410. doi:[10.1109/TNET.2002.1012370](https://doi.org/10.1109/TNET.2002.1012370)
33. Gustafsson E, Jonsson A, Perkins CE (2004) Mobile IPv4 regional registration, draft-ietf-mobileip-reg-tunnel-09.txt
34. Soliman H, Castelluccia C, El Malki K, Bellier L (2005) Hierarchical mobile IPv6 mobility management (HMIPv6), IETF RFC 4140
35. Kempf J (2001) Dormant mode host alerting (IP paging) problem statement. IETF RFC 3132
36. Kempf J, Castelluccia C, Mutaf P, Nakajima N, Ohba Y, Ramjee R, Saifullah Y, Sarikaya B, Xu X (2001) Requirements and functional architecture for an IP host alerting protocol. RFC 3154
37. Zhang X, Castellanos JG, Campbell AT (2002) P-MIP: paging extension for mobile IP. ACM Mobile Netw Appl 7(2):127–141. doi:[10.1023/A:1013774805067](https://doi.org/10.1023/A:1013774805067)
38. Hong K, Lee SK (2012) Dynamic multi-step paging scheme in PMIPv6-based wireless networks. Wireless Netw 18:33–44 [PPMIP] doi:[10.1007/s11276-011-0385-0](https://doi.org/10.1007/s11276-011-0385-0)
39. El Malki K (ed) (2007) Low-latency handoffs in mobile IPv4. RFC 4881, June 2007
40. Koodli R, Perkins C (2007) Mobile IPv4 fast handovers. RFC 4988, October 2007
41. Koodli R (ed) (2009) Fast handovers for mobile IPv6, IETF RFC 5568
42. IETF RFC 4861: Neighbor discovery for IP version 6 (IPv6)
43. Yokota H, Chowdhury K, Koodli R, Patil B, Xia F (2010) Fast handovers for proxy mobile IPv6, IETF RFC 5949 [FPMIPv6]

44. Ryu S, Mun Y (2009) Scheme to prevent packet loss during PMIPv6 handover. *Proc Trans Comput Sci* VI 5730:131 (PL-PMIPV6)
45. Bargh MS, Hulsebosch B et al (2008) Reducing handover latency in future IP-based wireless networks: proxy mobile IPv6 with simultaneous bindings. In *Proceedings of international symposium world wireless, mobile multimedia network*, Newport Beach, CA, USA, 2008, pp 1–10 (PMIP-BI)
46. Kim JI, Koh SJ (2011) Proxy mobile IPv6 with partial bicasting for seamless handover in wireless networks. In *Proceedings of ICOIN*, Barcelona, Spain, 26–28 Jan 2011, pp 352–356 (PMIP-BIP)
47. Kent S, Seo K (2005) Security architecture for the internet protocol. IETF RFC 4301
48. Kaufman C, Hoffman P, Nir Y, Eronen P (2010) Internet key exchange protocol version 2 (IKEv2), IETF RFC 4718
49. Eronen P (2006) IKEv2 mobility and multihoming protocol (MOBIKE), IETF RFC 4555
50. Rigney C, Willens S, Rubens A, Simpson W (2000) Remote authentication dial in user service (RADIUS), IETF RFC 2865
51. Fajardo V, Arkko J, Loughney J, Zorn G (2012) Diameter base protocol. IETF RFC 6733
52. Glass S, Hiller T, Jacobs S, Perkins C (2000) Mobile IP authentication, authorization, and accounting requirements. IETF RFC 2977
53. Calhoun P, Johansson T, Perkins C, Hiller T, McCann P (2005) Diameter mobile IPv4 application. IETF RFC 4004
54. Chowdhury K, Lior A, Tschofenig H (2006) RADIUS mobile IPv6 support, draft-chowdhury-mip6-radius-02.txt
55. Xia F, Sarikaya B, Korhonen J, Gundavelli S, Damic D (2012) RADIUS support for proxy mobile IPv6. IETF RFC 6572
56. Korhonen J, Bournelle J, Muhanna A, Chowdhury K, Meyer U (2008) Diameter proxy mobile IPv6: support for mobile access gateway and local mobility anchor to diameter server interaction. draft-korhonen-dime-pmip6-04.txt

Chapter 7

Mobility Management at Transport Layer

This chapter is dedicated to transport-layer mobility support schemes, which follow an end-to-end philosophy, putting the notion of mobility at the end nodes, and without any requirements on network infrastructure devices such as routers and servers. The existing schemes will be introduced, along with the explanations of their technical principles. The mobile Stream Control Transmission Protocol (mSCTP) is used as a typical example of mobility management at the transport layer and is covered with a detailed introduction and in-depth analysis.

7.1 Overview

The logic of mobility management at the transport layer derives from the most widely used transport-layer protocol—Transmission Control Protocol (TCP). In general, a TCP connection is identified by a quad-tuple: <the source IP address, the source port number, the destination IP address, the destination port number>. When the mobile node (MN) changes its attachment point because of mobility, its IP address will change accordingly. The existing TCP connection will therefore be terminated. The goal of transport-layer mobility management was to maintain the existing connection when the MN changes IP addresses during movement, thus eliminating the connection re-establishment operation.

Mobility management at the transport layer follows an end-to-end philosophy, which places the notion of mobility at the end node, without imposing any requirements on network infrastructure devices such as routers and servers in most schemes. Various transport-layer mobility management schemes have been proposed. Based on the different methods and functions used, these schemes can be classified into four types [1]:

- Gateway-based mobility scheme: A special gateway (or proxy) is introduced to maintain continuous communications.

- Connection migration protocol: MN notifies the correspondent node (CN) about the change of IP addresses at the MN and migrates the connection from the old address to the new address.
- Transport-layer handover protocol: It provides a handover function together with performance considerations on handover latency, data loss, etc.
- Complete mobility management scheme: complete mobility support with both handover and location management.

The above scheme types all have their different mobility support capabilities, functions, performances, and technical principles. Further details, analysis, and comparisons of these scheme types, along with introductions of some example schemes, are given in Sect. 7.2.

7.2 Transport-Layer Mobility Support Solutions

7.2.1 Gateway-Based Mobility Scheme

A special gateway (or proxy) is introduced in gateway-based mobility schemes to maintain continuous communication between the CN and the MN. This gateway is often placed at an intermediate point between the CN and the MN, and the original end-to-end connection is split into two portions: the connection between the CN and the gateway, and the connection between the gateway and the MN. Figure 7.1 illustrates the technical principles of the gateway-based mobility scheme. The CN-gateway connection will remain unchanged, while the gateway-MN connection may change, depending on the availability status of the MN's multiple interfaces. The gateway is the essential entity for mobility support in this scenario. It is responsible for connection splitting, for transferring or redirecting data from the old gateway-MN connection to the new one, and also for location management in some schemes.

Because the intermediaries (i.e., the gateway or proxy) are introduced to monitor the TCP traffic and actively participate in the flow control to enhance performance,

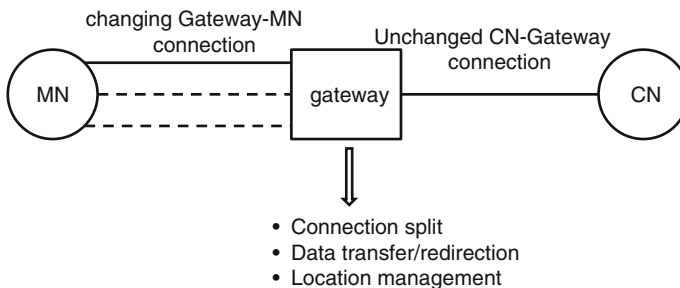


Fig. 7.1 Technical principles of gateway-based mobility scheme

gateway-based mobility schemes only simulate the end-to-end semantics but do not offer true end-to-end signaling. As a result, these techniques are not applicable when the IP payload is encrypted [2].

The CN-gateway connection is often a TCP connection, but it is not necessary to use TCP over the link between the gateway and the MN. One can use any other protocol optimized for wireless links; for example, Mobile TCP (M-TCP) uses a selective repeat protocol (SRP) over the wireless link [2].

MSOCKS, Indirect TCP (I-TCP), Mobile TCP (M-TCP), Mobile User Datagram Protocol (M-UDP), and Bay Area Research Wireless Access Network (BARWAN) are typical examples of the gateway-based mobility schemes.

(1) MSOCKS

MSOCKS [3] uses the TCP splice proposed in [4] to split a TCP connection at a proxy. The host-to-host communication is thus divided into CN-proxy communication and proxy-MN communication. The CN-proxy connection remains unchanged, while the proxy-MN connection may be interrupted because of the unavailability of the original IP address. The MN will disconnect itself from the old subnet, acquire a new IP address at another interface from the new subnet, and then establish a new connection with the gateway. The gateway is responsible for transferring the data from the old proxy-MN connection to the new connection.

The proxy in MSOCKS is also responsible for the location management function, which records and maintains the location changes of the MNs.

The advantage of MSOCKS is that it limits the mobility within the coverage range of the proxy alone.

(2) I-TCP and M-TCP

I-TCP [5] introduces mobility support routers (MSRs) as intermediate gateways between the CN and the MN. A CN-gateway TCP connection and a gateway-MN I-TCP connection over a wireless link are established to provide CN to MN communication. When the MN moves, the CN-gateway TCP connection remains unchanged as a new connection between the MN and the gateway is established to replace the old connection. M-TCP [6] is an enhanced version of I-TCP with lower complexity in the wireless part of the connection.

I-TCP and M-TCP do not support IP diversity and soft handover. They also do not include location management functions.

(3) M-UDP

M-UDP [7] is a mobility support scheme based on UDP. Similar to I-TCP and M-TCP, a gateway is introduced for mobility support following the connection split method. Like UDP, M-UDP does not guarantee the reliable delivery of the datagrams. However, unlike UDP, it does ensure that the number of lost datagrams remains small.

Likewise, M-UDP does not support IP diversity and soft handover.

(4) BARWAN

BARWAN [8] is a solution proposed based on the heterogeneous wireless overlay network environment. It has a gateway-centric architecture based on the assumption that the wireless networks are built around the gateways. Software is required at the MN and at the gateway to support mobility. BARWAN supports IP diversity and therefore also supports soft handover. Because BARWAN defines the mobility decision function at the upper-layer application, the mobility is not transparent to the upper layer. In addition, BARWAN does not provide the location management function.

7.2.2 *Connection Migration Protocol*

Connection migration is one possible scheme that could be used to ensure that the CN could maintain communications with the MN when the MN changes its IP address. This involves notifying the CN about the change of IP address at the MN and migration of the connection from the old address to the new address. However, the connection migration scheme does not handle handover issue, resulting in a probable temporary stop in the data flow during the migration process.

Typical connection migration protocols include Freeze-TCP and TCP redirection (TCP-R), which are the extensions of TCP that allow a connection to be stopped and restarted during communication.

(1) Freeze-TCP

Freeze-TCP [2] was proposed based on the drawbacks of certain gateway-based schemes and an in-depth analysis of the transmission control mechanism of TCP. In Freeze-TCP, the MN advertises a zero window size to the CN to stop (i.e., freeze) the existing TCP connection during handover when it senses an impending disconnection and unfreezes the connection after the handover. This scheme can reduce packet losses during handover at the cost of higher latency. It also enhances the TCP throughput in the presence of frequent disconnections. Freeze-TCP is said to be a true end-to-end scheme and does not require the involvement of any intermediate nodes (e.g., gateways or proxies). In addition, it does not require any changes to the TCP code on the CN side. Changes to the TCP code are confined entirely to the MN side and thus guarantee interoperability with the existing infrastructure.

Freeze-TCP only handles connection migration, but without handover or location management functions. It can, however, be integrated with other schemes to provide complete mobility management.

(2) TCP-R

TCP-R [9] is a connection migration scheme that maintains active TCP connections when the disconnection occurs because of a change of IP address or a change of the network device.

TCP-R handles the existing active connections during the handover using a straightforward mechanism of updating the end-to-end address pairs in the TCP connections. It assumes that the MN could detect the change of IP address in some way. Then, the MN issues the redirection message to inform the CN of its new IP address. When the message is received by the CN, the pair of addresses in the existing TCP connections is modified using the message, and they then resume communication with each other through the revised TCP connections. Thus, TCP-R can maintain continuous operation with minimal overheads and complexity.

To establish the new MN connection, mobility supporting mechanisms at other layers, such as Mobile IP (MIP) or dynamic domain name system (DDNS), can be integrated with TCP-R to provide a compensative operation.

TCP-R provides mobility support that is transparent to the upper applications. However, the TCP code at both the MN and the CN must be modified. Also, it relies on MIP or DDNS for location management functions.

7.2.3 Transport-Layer Handover Protocol

The transport-layer handover protocols are the extensions of the traditional transport-layer protocols that aim to provide a handover function together with performance considerations, such as to reduce handover latency and data loss. They can only support handover, without any location management function.

Radial reception control protocol (R^2CP), mobile multimedia streaming protocol (MMSP), and mSCTP could be used as typical examples of transport-layer handover protocols.

(1) R^2CP

The R^2CP [10] was designed based on the reception control protocol (RCP). Focusing on a scenario where the mobile host (MH) acts as the receiver, RCP acts as a TCP clone in terms of its general behavior, but moves the congestion control and reliability control from the sender side to the receiver side for improved congestion control, loss recovery, and power management. R^2CP was proposed with the additional consideration of the MHs being equipped with heterogeneous wireless interfaces. Connection management, packet scheduling, single interface-based congestion control, and aggregate connection-based flow control functions are designed in R^2CP to support seamless handover, service migration, and bandwidth aggregation.

R^2CP has no location management function. However, other location management schemes may be integrated with R^2CP to provide a complete mobility management solution.

(2) MMSP

The MMSP [11] was designed as an end-to-end robust IP soft-handover protocol at the transport layer. It was proposed to be capable of multi-homing and bicasting

in combination with forward error correction (FEC) to address the performance problems of the existing mobility management protocols.

To eliminate packet losses during handover, MMSP uses a packet path diversity scheme and develops an end-to-end bicasting mechanism that enables IP soft handover. MMSP also uses an FEC scheme embedded in the bicasting mechanism and fragmentation to offset the wireless errors.

In addition, MMSP does not provide the location management function.

(3) mSCTP

mSCTP [12] is an extension of Stream Control Transmission Protocol (SCTP) with dynamic address reconfiguration (DAR) functions. mSCTP can provide end-to-end mobility support at the transport layer because of its multi-homing feature and its DAR extension. mSCTP supports IP diversity through its multi-homing feature. Multiple IP addresses can be configured at the mSCTP endpoint, and one of them is set as the primary address used for data transmission. The DAR extension defines the solution by adding a new IP address, changing the primary IP address, and deleting the old IP address dynamically for soft handover support.

Similar to the previous handover protocols, mSCTP does not provide a location management function. Other location management schemes such as DDNS, MIP, and SIP can be integrated with mSCTP for the location management function.

7.2.4 Complete Mobility Management Schemes

Complete mobility management schemes provide complete end-to-end mobility management at the transport layer, including handover and location management functions. Migrate and seamless IP diversity-based generalized mobility architecture (SIGMA) is a typical example.

(1) Migrate

Migrate [13] is another extension to TCP for mobility support. Three important components are designed as part of the migrate system: addressing, mobile host location, and connection migration. The system uses secure updates to the DNS following an address change to allow Internet hosts to locate a mobile host, and a set of connection migration options are used to securely and efficiently negotiate a change in the IP address of a peer without breaking the end-to-end connection.

However, migrate requires modification to the transport-layer protocol for the communication endpoints.

(2) SIGMA

SIGMA [14] is a complete mobility management scheme at the transport layer. The basic idea of SIGMA is to exploit the IP diversity to keep the old path alive during the process of setting up the new path to achieve seamless handover with low

latency and low packet loss. The soft handover in SIGMA follows the SCTP-based handover procedure.

SIGMA uses the DNS to provide the location management function.

7.2.5
Comparison

(1) Comparison of the different transport-layer mobility support types

A comparison of the different transport-layer mobility support scheme types is given in Table 7.1 in terms of technical principles, and their advantages and disadvantages, and examples are also given.

Table 7.1 Comparison of the different transport-layer mobility support types in terms of technical principles

Class	Description	Advantages	Disadvantages	Examples
Gateway-based mobility scheme	A special gateway (or proxy) is introduced to maintain continuous communication	• Mobility is limited only at the gateway-MN connection portion, because the CN-gateway connection remains unchanged	• Infrastructure change requirements • Not true end-to-end signaling • Single point failure problem	MSOCKS I-TCP M-TCP M-UDP BARWAN
Connection migration protocol	Notifies the CN about the change of IP address at the MN and migrates the connection from the old address to the new address	• The communication continues with the connection migration function	• No handover function, which may result in a temporary stop in the data flow during migration	Freeze-TCP TCP-R
Transport-layer handover protocol	Provides handover function together with performance consideration of handover latency and data loss.	• Supports handover with performance ensured to a certain extent	• No location management function	R ² CP MMSP mSCTP
Complete mobility management scheme	Complete mobility support with handover and location management	• Complete mobility management functions and capabilities	–	Migrate SIGMA

Table 7.2 Comparison of the different transport-layer mobility support types in terms of their functions [1]

Criteria	Gateway-based mobility scheme	Connection migration protocols	Handover protocols	Complete mobility management schemes
Handover	Usually hard, but soft for BARWAN	N/A	Soft	Soft, but hard for migrate
Fault tolerance	Fails if the gateway fails	Yes	Yes	New connections would fail if the location management server fails
Transparency	Usually yes, but no for BARWAN	Yes at CN, no at MN	Yes	Yes, but no for migrate
Loss/delay	On the fly packets are lost, no for BARWAN	On the fly packets are lost for R-TCP, but Freeze-TCP prevents losses	No	No, migrate stops transmission
Conflicts with security solutions	Yes	No	No	No
IP diversity	No, supported for BARWAN	No	Supported	Supported, but not for migrate
Change in infrastructure	Yes	No	No	No
Change in protocol stack	Yes	Yes	Yes	Yes

Table 7.2 gives a more detailed comparison between the above types in terms of their mobility-related functions.

(2) Comparison of the different transport-layer mobility support schemes

Table 7.3 provides a detailed comparison of the above mobility support schemes at the transport layer in terms of their functions.

7.3 Typical Protocol: mSCTP

7.3.1 Overview

mSCTP is an extension of the SCTP protocol with mobility support capability.

SCTP is the transport-layer protocol defined by the IETF in RFC2960 [15], which was published in October 2000. The primary purpose of SCTP was to

Table 7.3 Comparison of the different transport-layer mobility support schemes [1]

Scheme	Handover	Transparency	IP diversity	Change in infrastructure	Change in protocol stack		Location management
					MN	CN	
MSOCKS	Hard	Yes	Supports	Yes	Yes	No	Supports
SIGMA	Soft	Yes	Supports	No	Yes	Yes	Supports
Migrate	Hard	Yes	No	No	Yes	Yes	Supports
Freeze-TCP	–	Yes	No	No	Yes	No	No
R ² CP	Soft	Yes	Supports	No	Yes	Yes	No
MMSP	Soft	Yes	Supports	No	Yes	Yes	No
I-TCP	Hard	Yes	No	Yes	Yes	No	No
M-TCP	Hard	Yes	No	Yes	Yes	No	No
M-UDP	Hard	Yes	No	Yes	Yes	No	No
BARWAN	Soft	No	Supports	Yes	Yes	Yes	No
TCP-R	–	Yes	No	No	Yes	Yes	Supports
mSCTP	Soft	Yes	Supports	No	Yes	Yes	No

provide a reliable end-to-end message transportation service over IP-based networks, and it was then developed as a general-purpose transport protocol for data communications. SCTP resides at the same location in the IP architecture as TCP and UDP, which uses the service provided by the lower-layer connectionless packet network and provides a transmission service to the upper-layer applications. SCTP has some similar features to those of TCP, e.g., connection-oriented, end-to-end duplex transmission, and reliability. It also has some new features, such as multi-homing and multi-streaming.

mSCTP extended SCTP by using a dynamic address reconfiguration (DAR) extension. It integrated the DAR extension and the multi-homing feature for the end-to-end mobility support capability at the transport layer.

In this section, the SCTP and mSCTP features that are helpful for understanding mobility management functions of mSCTP are introduced.

(1) SCTP endpoint and association

mSCTP follows the basic definitions of the SCTP endpoint and association. In SCTP, the communication parties are called endpoints, and the communication relationship between them is called an association. The SCTP transport address is used as the unique identifier of each SCTP endpoint. The SCTP transport address is defined as a combination of an IP address (or an IP address set, depending on whether the endpoint is multi-homed or not) and an SCTP port. For example, the SCTP endpoint on machine B shown in Fig. 7.2 can be denoted as [114.255.40.211: 5432], where 114.255.40.211 is the IP address of the network interface (NI) of host B.

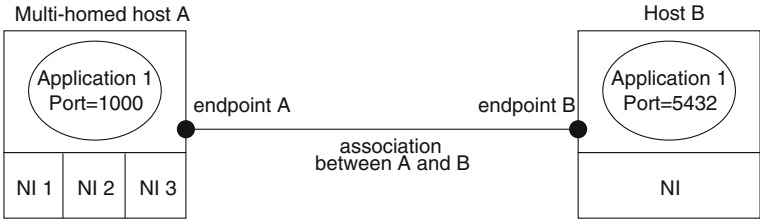


Fig. 7.2 Sctp endpoint and association

As mentioned above, Sctp supports the multi-homing feature, which means that an Sctp endpoint can have multiple IP addresses. This type of Sctp endpoint is called a multi-homed Sctp endpoint. The multi-homed Sctp endpoint can be represented by a list of Sctp transport addresses on the host that shares a single Sctp port. For example, the Sctp endpoint serving Application 1 on host A in Fig. 7.2 can be denoted by:

$$\text{endpoint} = [59.64.139.20, 211.68.71.30, 10.108.100.125: 1000]$$

where 59.64.139.20, 211.68.71.30, and 10.108.100.125 are the IP addresses of NI 1, NI 2, and NI 3, respectively, on host A.

Sctp association is the communication relationship between the Sctp endpoints. It can be conveniently identified using the Sctp endpoint pair. For example, the association between machine A and machine B above can be denoted by:

$$\text{Association} = \{[59.64.139.20, 211.68.71.30, 10.108.100.125: 1000] : [114.255.40.211: 5432]\}$$

Sctp is a connection-oriented transport protocol, as mentioned earlier. This means that before any application data can be transported from one Sctp endpoint to another, the two Sctp endpoints must go through a setup procedure to establish the Sctp association by exchanging state information. Only the basic concepts that are helpful in understanding the Sctp protocol are presented here. For full details of the Sctp protocol, please refer to [15, 16].

(2) Multi-homing

In general, multi-homing may be the result of the following two scenarios: (1) A host is configured with multiple network interface cards, with each being assigned to a different IP address; (2) a host is configured with only a single network interface card, but multiple IP addresses have been assigned. This is the general scenario in IPv6 networks, where a network interface card will be assigned to a global address, a site-local address, and a link-local address.

Specifically, multi-homing in Sctp refers to an Sctp endpoint that has multiple IP addresses used for data transmission. Therefore, we only discuss the scenario in

which multiple interfaces are installed, and each interface has only one IP address within the context of this chapter.

The primary purpose of the multi-homing feature in SCTP is to provide an error tolerance capability at the network level. Considering a multi-homed host as the data receiver, each IP address of the host represents a data path from the data sender side to the receiver side. Thus, the sender can select one path, called the primary path, for data transmission. The other available paths are defined as secondary paths. When the primary path becomes unavailable because of interface failure or network congestion, SCTP can handover the data transmission to another path and thus improves the error tolerance capability of the SCTP association.

In the SCTP extension scheme [17], the multi-homing feature is also used for load balancing. All available paths are used for data transmission concurrently to improve the throughput of the SCTP association.

Also, multi-homing is one of the two important features of mSCTP for mobility support. It is integrated with the DAR extension for end-to-end mobility at the transport layer.

(3) DAR extension

The DAR extension was defined in [18] by the IETF, and it provides an IP address reconfiguration mechanism for an active association. Here, the IP address reconfiguration operations include addition of a new IP address, deletion of an old IP address, and changing the primary IP address.

The DAR extension introduces two new control chunk types, address configuration change chunk (ASCONF) and address configuration acknowledgment chunk (ASCONF-ACK), together with the required parameter types and error reasons.

A sender of an ASCONF message uses the parameters in the ASCONF control chunk to indicate the operations. The major parameters related to dynamic address reconfiguration are *Add IP Address*, *Delete IP Address*, and *Set Primary IP Address*.

- *Add IP Address* parameter is used to tell the correspondent endpoint to add new IP address(es) to the existing association. The *Address Parameter* field in this parameter is used to carry the specific IP address(es), which can be a single address or multiple addresses.
- *Delete IP Address* parameter is used to tell the correspondent endpoint to delete IP address(es) from the existing association. The *Address Parameter* field is again used to carry the single or multiple addresses to be deleted.
- *Set Primary IP Address* parameter is used to tell the correspondent endpoint to change the primary IP address of the existing association. The *Address Parameter* field is used to indicate the new primary address. The *Set Primary IP Address* parameter can also be used in the INIT chunk and the INIT-ACK chunk in the establishment of the initial association to indicate the original preference for the primary address.

The communication endpoints exchange the ASCONF control chunk and the ASCONF-ACK control chunk for dynamic address reconfiguration.

- The ASCONF control chunk carries the ASCONF parameter, which may be the aforementioned *Add IP Address* parameter, *Delete IP Address* parameter, or *Set Primary IP Address* parameter, to indicate the required address reconfiguration operations. In addition, multiple address reconfiguration parameters can be included in a single ASCONF control chunk.
- The ASCONF-ACK control chunk is the acknowledgment for the ASCONF chunk. It defines the *ASCONF Parameter Response* field to indicate the processing state of the corresponding ASCONF chunk. If there are any errors, the error reasons will be indicated in the *Error Cause Indication* field.

For more details about the DAR extension and the related parameters, please refer to [18].

(4) Mobility support based on mSCTP

The mobility support for mSCTP is realized based on the multi-homing feature and the DAR extension. This mobility support is based on the end-to-end semantics and does not place any requirements on the network infrastructure. The IETF and many academic researchers have devoted lots of efforts to the protocol procedures, handover performance analysis and improvement, and integrated mobility management function design.

For a client–server session initiated by a mobile host to a fixed host, the handover based on mSCTP can be described as follows. This handover procedure follows the soft-handover concept to reduce packet losses during handover.

- The mobile host acquires a new IP address from the access network that it is approaching;
- The mobile host adds this new IP address to the existing active association;
- The mobile host changes the primary IP address of the existing association;
- The mobile host deletes the unavailable IP address from the existing association.

It should be noted that mSCTP-based mobility is limited only to the above type of client–server session that is initiated by the mobile host to the fixed host. For communications initiated by a fixed host to a mobile host, or for peer-to-peer sessions, mSCTP can only provide the handover function. The mSCTP should be integrated with other technologies such as MIP, SIP, DDNS, or reliable server pooling (RSerPool) for location management functions.

A detailed introduction and analysis based on the mobility management reference model will be given in the following sections.

7.3.2 Location Management

As mentioned above, mSCTP does not define the location database in the protocol architecture and has no location management function. mSCTP can be used in conjunction with other technologies such as MIP, SIP, DDNS, or RSerPool for location management.

The IETF discussed location management for mSCTP in [19]. For sessions that are terminated at the mobile host, several possible location management mechanisms were proposed to locate the mobile host.

(1) Mobile IP-based location management

The details of the mobility management functions of Mobile IP were introduced in Chap. 6. Specifically, in the mSCTP/Mobile IP integration scenario, MIP is only used for the location management function in the sessions initiated by the fixed CN to the MN. Based on the location management function of MIP, the CN is enabled to locate the MN and establish the SCTP association. Once the association has been successfully established, the subsequent data transmission is conducted based on mSCTP over IP, and the on-going SCTP session is supported by the mSCTP soft-handover function.

The location management function of MIP is illustrated in Fig. 7.3. Figure 7.3a shows the standard SCTP association establishment defined in RFC2960 [15]. The communicating parties will exchange INIT, INIT-ACK, COOKIE-ECHO, and COOKIE-ACK to establish the association based on the four-handshake procedure.

When the MIP location management function is used, the HA is introduced to accomplish the association establishment, as shown in Fig. 7.3b. The key point for association establishment of sessions that terminate at the MN is how the first INIT control chunk sent by the fixed CN can arrive at the MN away from its original

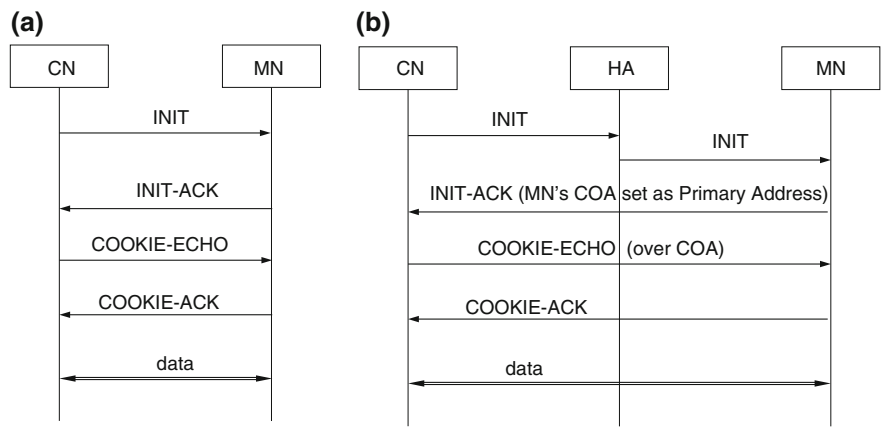


Fig. 7.3 mSCTP association establishment based on the location management function provided by MIP. a standard SCTP [15]. b mSCTP over Mobile IP [19]

location. When considering the location management function of MIP, the HA is responsible for maintaining current location information for the MNs. When an MN moves into the visiting network, it will update its location information in the HA via a binding update procedure. Therefore, the first INIT chunk can be forwarded by the HA to the MN. When the MN receives the INIT chunk, it can respond with the INIT-ACK control chunk directly to the CN, without forwarding by the HA. This INIT-ACK chunk must contain the care-of address (CoA), which can be addressable to the current location of the MN. In addition, this CoA will be set as the primary address of the responding association.

After the association has been established, data transmission between the CN and the MN relies on mSCTP over IP. Tunneling between the HA and the MN is not used. Also, the home address (HoA) of the MN is not involved in the data transmission and is only used for location management.

If the MN moves continuously during the data transmission, the handover will be handled by mSCTP as described in Sect. 7.3.3. The handover function of MIP will thus not be used here.

To summarize, MIP is only used for location management, and particularly for locating the MN in the association establishment procedure in this scenario.

(2) SIP-based location management

The details of the mobility management functions of SIP will be introduced in Chap. 8. For location management in particular, the location server is responsible for maintaining the location information of all users. When the MH moves into a visiting network, it will send a SIP REGISTER message to its home SIP registrar server, which will then update the location server accordingly.

When the location management function of SIP is applied to mSCTP, these two protocols can easily be used together. This is because SIP is an application-layer protocol, while mSCTP acts at the transport layer. The host uses mSCTP instead of TCP/UDP as the transport-layer protocol to integrate these two protocols.

In such a scenario, SIP is responsible for both location management and call setup. The location management entities and operations will thus be applied. For a call setup requested with a MH, the home SIP proxy server will interrogate the location database to locate the MH and then relay the SIP INVITE message to the current SIP proxy server and up to the MH. These operations will result in successful SCTP association establishment via SIP signaling. After that, the data transmission will be performed based on mSCTP. When the MH moves to a new location, it will also be mSCTP that is responsible for the handover function.

In summary, SIP entities (i.e., the location server and the registrar) and register operations are used in such scenarios for location management. The call setup for communications that terminate at the MH will be completed based on SIP signaling.

(3) RSerPool-based location management

RSerPool [20] can be used for location management. A mobile server (MS) registers a pool handle such that it becomes part of a pool. It is permissible for

a pool to consist of one pool element only. A client (whether mobile or not) must know the pool handle of the mobile server to be able to talk to it. The client sends a name resolution request to one of the endpoint handlespace redundancy protocol (ENRP) servers and receives the current IP addresses in return. Because the ENRP servers within their operational scope share their states, it is not important which ENRP server is contacted. If the MS changes its IP address, it then re-registers at the home ENRP server.

The pool handles can therefore be used to address a server with changing IP addresses. If the mobile client (MC) or the MS change their addresses because of handovers, then mSCTP can be used to handle this situation, except in the case where the MC and the MS change their addresses simultaneously. In this case, mSCTP fails, i.e., the SCTP association terminates. The RSerPool session concept can be used to re-establish a new SCTP association based on the new addresses and continues the RSerPool session. Depending on the application, the effect of this session failover on the application can be very small.

7.3.3 Handover Control

Handover control in mSCTP relies on the aforementioned multi-homing feature and the DAR extension. The basic handover procedure, the handover control functions, and the research issues for mSCTP-based handover are introduced in this section.

(1) Basic handover procedure

mSCTP provides a handover control function based on its multi-homing feature and DAR extension. Figure 7.4 shows an example scenario for mSCTP-based handover. In Fig. 7.4, MN denotes the multi-homed hosts with multiple network interfaces, which are represented by NI1 and NI2. The CN is a fixed host. The communication is initiated by the MN and terminated at the CN. When the MN resides within the coverage area of access router AR1, it connects to the network via network interface IN1, with IP1 as its address. With the movement of the MN along the direction indicated by the arrow in Fig. 7.4, the MN will travel through the coverage area of AR1 and enter the area covered by AR2. A new address, IP2, will then be acquired from NI2, through which the MN will connect to the network. The following discussion of mSCTP-based handover is based on this simple but typical scenario.

A. mSCTP-based handover procedure

Considering the scenario shown in Fig. 7.4, *Location1* and *Location2* represent the MN locations in AR1 and AR2, respectively. The MN is assumed to initiate the communication to the CN at *Location1* and establishes the SCTP association successfully. At this time, the IP address of the MN is IP1. With the movement of the MN along the direction indicated by the arrow shown in Fig. 7.4, the MN

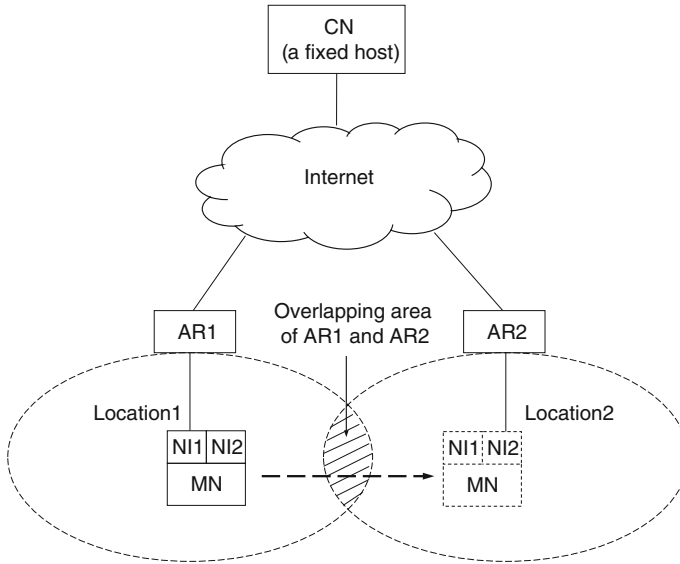


Fig. 7.4 Sample scenario of mSCTP-based handover control

moves to *Location2*. A handover operation is thus necessary to maintain communication continuity. The handover procedure can be divided into the following four stages.

- a. Acquisition of a new IP address in the AR2 coverage area

During the movement of the MN from *Location1* to *Location2*, it will enter the overlap area of AR1 and AR2 indicated by the hatched section in Fig. 7.4. The MN will be assigned to a new address, IP2, from AR2 through the Dynamic Host Configuration Protocol (DHCP) or by the stateless address configuration in IPv6.

- b. Adding the new IP address to the SCTP association

After acquisition of the new IP address, the MN will send the ASCONF message carrying the *Add IP Address* parameter (introduced in Sect. 7.3.1) to tell the CN to add this new address to the existing association. Then, the MN will receive the corresponding acknowledgment message, ASCONF-ACK, from the CN.

- c. Changing of the primary IP address

The MN moves continuously toward *Location2*. The data transmission performance through AR2 becomes better than that through AR1 (this is because AR1 will gradually become unavailable in this example), and thus, IP2 should be set as the primary address for data transmission. The MN will send the ASCONF message carrying the *Set Primary IP Address* parameter to notify the CN. After the MN receives the corresponding acknowledgment message, it begins to transport data through IP2.

d. Deleting the old IP address from the association

When the MN moves continuously toward *Location2*, it will leave the coverage area of AR1, and the old IP address (i.e., IP1) will become unavailable. At this time, the MN will send an ASCONF message carrying the *Delete IP Address* parameter to tell the CN to delete IP1 from the association. Then, the MN will receive the acknowledgment message from the CN.

B. mSCTP-based handover signaling

Based on the handover procedure described above, the handover signaling process can be described as shown in Fig. 7.5.

(2) Handover control functions

In Chap. 2, we mentioned the major control functions in the handover process: the handover rule, the handover control mode, resource allocation during handover, and communication link transfer. These control functions in mSCTP-based handover will be introduced in this section.

A. Handover rule

In mSCTP-based handover, the handover triggering operation involves one communication party sending an ASCONF message with the *Set Primary IP*

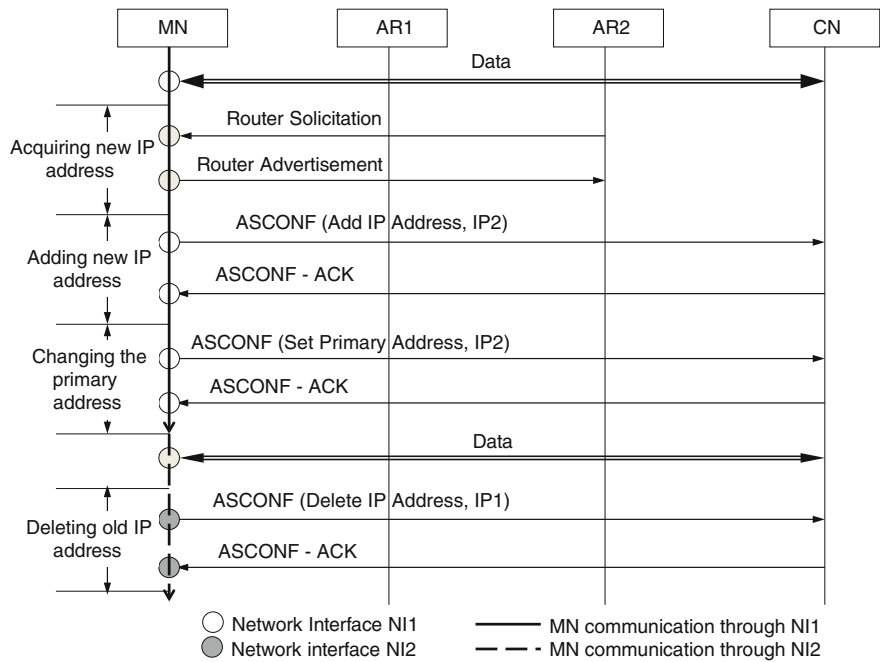


Fig. 7.5 mSCTP-based handover signaling

Address parameter to modify the primary IP address of the SCTP association. For a multi-homed SCTP endpoint with multiple interfaces and thus with multiple IP addresses, the inherent handover rule determines the handover triggering criteria and time. This rule can be extended to determination of the time to add the new IP address to the association and the time to delete the old IP address from the association.

These handover rules are not defined explicitly in the mSCTP protocol standards. However [19] suggests some possible handover rules for mSCTP. In addition, the handover rules, together with some necessary functionality extensions, have also been proposed in some existing research papers [21–26]. It should be noted that the handover rules for mSCTP are often integrated into the handover decision in vertical handover control.

The possible handover rules for mSCTP with the necessary functionality extensions are summarized in Table 7.4.

B. Handover control mode

We introduced the commonly used handover control modes, network-controlled handover (NCHO), mobile-controlled handover (MCHO), and mobile-assisted handover (MAHO), in Sect. 2.3.3. Because mSCTP provides end-to-end mobility support without involving any network infrastructure devices, the handover control functions should be implemented at the MN. Therefore, MCHO is the most suitable handover control mode for mSCTP-based handover.

C. Resource allocation during handover

In mSCTP-based handover control, the related resource allocation during handover is mainly IP address acquisition and network attachment to the new access network. For example, the allocation may be IP address assignment based on DHCP in a WLAN. Another example is the GPRS attachment and packet data protocol (PDP) context activation procedure.

D. Communication link transfer

Communication link transfer is executed at the transport layer. The data transmission will be transferred to the new path when the MH sends the ASCONF message with the *Set Primary Address* parameter to trigger the handover.

(3) mSCTP extensions for vertical handover

As described in Sect. 3.1.2, vertical handover is the handover between heterogeneous access technologies. mSCTP supports end-to-end handover and resides in the transport layer, which is unrelated to the access technologies. Therefore, it is used for vertical handover in many schemes.

As we know, mSCTP is an extension of SCTP. However, SCTP was originally proposed for wired networks. If mSCTP is to be directly applied to overlapped heterogeneous wireless networks for vertical handover, new features should be

Table 7.4 Handover rules for mSCTP with necessary functionality extensions

Handover rules	Description	Related factors	Necessary functionality extensions
The primary IP address is updated as soon as a new IP address is assigned	As soon as the MN receives the acknowledgment of the added IP address operation from the CN, it sends out another ASCONF message with the <i>Set Primary IP Address</i> parameter to change the primary address of the association, and thus triggers the handover operation	No	No
The handover is triggered by an explicit notification from the lower layers	The lower layers give an explicit handover notification to the mSCTP based on measurement and comparison at the different network interfaces. Accordingly, the MN sends the ASCONF message with the <i>Set Primary IP Address</i> parameter to trigger the handover operation	The received signal strength at each network interface	A cross-layer interaction interface between mSCTP and the lower layers
The handover is triggered by an instruction from the upper layer	This is often used in vertical handover scenarios where the MN is configured with multiple heterogeneous network interfaces that may be available simultaneously The upper layer of the MN is responsible for the multi-criterion handover decision and then sends the handover instruction to mSCTP. Accordingly, the MN sends the ASCONF message with the <i>Set Primary IP Address</i> parameter to trigger the handover operation	The bandwidth, delay, coverage, and service fees of the heterogeneous network interfaces	Measurement of the network status and performance based on a continuous HEARTBEAT message in mSCTP or other dedicated measuring methods

Table 7.5 Comparison of basic SCTP and mSCTP in vertical handover

	Basic SCTP	mSCTP in vertical handover
Intention of multi-homing	Used for error tolerance through link redundancy	Used for mobility support
Link types involved	Steady wired links	Unsteady wireless links with high error rates besides wired links
Link differences in handover	Can be ignored	Obvious link differences between heterogeneous wireless access technologies
Possible causes of data loss	Congestion Unreachable address	Congestion Unreachable address Handover Wireless link errors

noted for performance improvement. Table 7.5 gives a comparison of the basic SCTP and mSCTP in vertical handover applications.

Some researchers have evaluated mSCTP-based vertical handover performance through theoretical analysis, simulation, and prototype testing [27–30]. The performance is described in terms of the handover latency, the end-to-end throughput in a continuous handover scenario, and the packet losses incurred during handover. This is also an efficient method to determine valuable protocol improvement points.

Based on the above analysis of its technical features and handover performance, mSCTP should be extended to satisfy the handover performance requirements in particular for vertical handover.

It is necessary to design related mechanisms to detect the variance between the links before and after handover and to differentiate among the data losses incurred because of congestion, because of wireless link errors and because of the handover process itself. The congestion control and avoidance mechanisms should be extended to improve the handover performance. Of course, these functions could not be implemented at the transport layer alone, but rely on information and support from the physical layer, the data link layer and the network layer. Therefore, the cross-layer concept is necessary for these extensions.

As mentioned above, most of the mSCTP-based handover rules, including the triggering events for dynamic addition and deletion of the addresses, are based on the received signal strength. In the vertical handover scenario with apparent link variance, it is unfeasible to rely on the signal strength alone. This problem should be integrated with the vertical handover decision for feasible and efficient mechanisms.

7.3.4 Security Mechanisms

The security mechanism of mSCTP is designed based on that of SCTP, together with necessary considerations for multi-homing, the multi-streaming features, and

the DAR extension. mSCTP uses Transport-Layer Security over SCTP (TLS/SCTP), SCTP over IPsec (SCTP/IPSec), and Secure SCTP (S-SCTP) to provide key data privacy, registration, and authentication functions, along with signaling message integrity to a certain extent.

(1) TLS/SCTP

Transport-Layer Security (TLS) is a security protocol defined by the IETF in RFC2246 [31], which provides both privacy and data integrity between two communicating applications. The protocol is composed of two layers: the TLS Record Protocol and the TLS Handshake Protocol. The TLS Record Protocol sits at the lowest level, which is layered on top of a reliable transport protocol (e.g., TCP). The TLS Record Protocol provides connection security that has two basic properties: It is both private and reliable. The TLS Record Protocol is used for the encapsulation of various higher level protocols. One such encapsulated protocol, the TLS Handshake Protocol, allows the server and the client to authenticate each other and allows them to negotiate an encryption algorithm and associated cryptographic keys before the application protocol transmits or receives its first byte of data.

TLS is mainly used over byte-oriented reliable transport protocols (e.g., TCP). RFC3436 [32] defines the TLS/SCTP security mechanism, along with the necessary extensions for its message-oriented nature, the multi-streaming feature, and the DAR extension.

TLS-based user data transmission is only used for bidirectional streams in SCTP. For a bidirectional stream, the TLS connection should be established through the TLS handshake process to provide data integrity and data authentication functions. TLS ciphersuites are used to provide data integrity and privacy. For the authentication function, mutual X.509-based authentication, or X.509 authentication at the server side and the challenge handshake authentication protocol (CHAP) at the client side, can be adopted.

The TLS/SCTP security mechanism still faces the following problems:

- TLS/SCTP does not support unordered delivery and partial reliability extensions, because TLS is designed for reliable and ordered transmission;
- TLS/SCTP does not provide security for unidirectional streams, because TLS connections require bidirectional communication;
- TLS/SCTP provides a user data integrity function, but without integrity protection for the signaling messages (i.e., the control chunks used in SCTP).
- TLS/SCTP faces scalability problems with increasing numbers of streams, because a single TLS connection is required for each bidirectional SCTP stream. The corresponding handshake signaling procedures will incur large overheads and suffer latency.

(2) SCTP/IPSec

IPSec is the security architecture defined by the IETF for IP networks, including the Authentication Header (AH) protocol, the Encapsulating Security Payload (ESP) protocol, Internet Key Exchange (IKE), and related algorithms for authentication and encryption. IPSec provides authentication, data integrity, and confidentiality for the IP layer.

RFC3554 [33] defines the SCTP/IPSec security mechanism. SCTP uses the IPSec security service transparently to provide security protection for the end-to-end transported user data and the SCTP control chunks. The AH protocol is used to provide data integrity and authentication. If necessary, the ESP protocol can be used for confidentiality support.

The disadvantage of SCTP/IPSec exists in the DAR scenario. When a new IP address is added to the existing SCTP association, a new cryptographic key exchange and negotiation procedure is required to establish a new security association for the newly assigned IP address.

(3) S-SCTP

Secure SCTP (S-SCTP) [34] was proposed to overcome the shortcomings of the above security mechanisms that use standard security protocols together with SCTP (i.e., TLS/SCTP and SCTP/IPSec). S-SCTP is the integration of the security functionality into SCTP, which then provides the security functionalities in the transport layer itself, without the need for any other security protocols, and it is compatible with the standard SCTP.

S-SCTP provides integrity and authentication support for SCTP-based user data and SCTP control chunks. To minimize the transmission overheads, the Hash Message Authentication Code (HMAC) algorithm is executed for the whole SCTP packet (including all the chunks and the common header) to provide an integrity function, together with the hash algorithms and the shared key. S-SCTP also provides privacy support for the user data and control messages. Also, the encryption operation can be conducted for the data of a single application or even for a single message.

In addition, S-SCTP defines different security levels to provide flexible differentiated security protection granularity for different application requirements to optimize the overheads, including the transmission overhead and the computation overhead.

To ensure compatibility with the standard SCTP protocol, S-SCTP defines some new chunks that are used for security session initialization and termination, the re-key handshake, encrypted data transmission, and authentication. Detailed definitions of these chunks can be found in [34].

References

1. Atiquzzaman M, Reaz AS (2005) Survey and classification of transport layer mobility management schemes. In: IEEE 16th international symposium on personal, indoor and mobile radio communications, 2005 (PIMRC 2005), pp 2109–2115. doi:[10.1109/PIMRC.2005.1651818](https://doi.org/10.1109/PIMRC.2005.1651818)
2. Goff T, Moronski J, Phatak DS, Gupta V (2000) Freeze-TCP: a true end-to-end TCP enhancement mechanism for mobile environments. In: INFOCOM 2000: proceedings of the 19th annual joint conference of the IEEE computer and communications societies, pp 1537–1545. doi:[10.1109/INFCOM.2000.832552](https://doi.org/10.1109/INFCOM.2000.832552)
3. Maltz DA, Bhagwat P (1998) MSOCKS: an architecture for transport layer mobility. In: INFOCOM '98: proceedings of the 17th annual joint conference of the IEEE computer and communications societies, pp 1037–1045. doi:[10.1109/INFCOM.1998.662913](https://doi.org/10.1109/INFCOM.1998.662913)
4. Maltz DA, Bhagwat P (1999) TCP Splice for application layer proxy performance. *J High Speed Netw* 8:225–240
5. Bakre A, Badrinath BR (1995) I-TCP: indirect TCP for mobile hosts. In: Proceedings of the 15th international conference on distributed computing systems, pp 136–143. doi:[10.1109/ICDCS.1995.500012](https://doi.org/10.1109/ICDCS.1995.500012)
6. Haas ZJ, Agrawal P (1997) Mobile-TCP: an asymmetric transport protocol design for mobile systems. 1997 IEEE international conference on communications (ICC '97), pp 1054–1058. doi:[10.1109/ICC.1997.610040](https://doi.org/10.1109/ICC.1997.610040)
7. Brown K, Singh S (1996) M-UDP: UDP for mobile cellular networks. *ACM SIGCOMM Comput Commun Rev* 26:60–78. doi:[10.1145/242896.242901](https://doi.org/10.1145/242896.242901)
8. Katz RH, Brewer EA, Amir E et al (1996) The Bay area research wireless access network (BARWAN) compcon '96: 'Technologies for the Information Superhighway' Digest of Papers, pp 15–20. doi:[10.1109/CMPCON.1996.501742](https://doi.org/10.1109/CMPCON.1996.501742)
9. Funato D, Yasuda K, Tokuda H (1997) TCP-R: TCP mobility support for continuous operation. In: Proceedings of the 1997 international conference on network protocols, pp 229–236. doi:[10.1109/ICNP.1997.643720](https://doi.org/10.1109/ICNP.1997.643720)
10. Kim KH, Zhu Y, Sivakumar R et al (2005) A receiver-centric transport protocol for mobile hosts with heterogeneous wireless interfaces. *Wireless Netw* 11(4):363–382. doi:[10.1007/s11276-005-1763-2](https://doi.org/10.1007/s11276-005-1763-2)
11. Matsuoka H, Yoshimura T, Ohya T (2003) End-to-end robust IP soft handover. In: IEEE international conference on communications, 2003 (ICC '03), 1:532–536. doi:[10.1109/ICC.2003.1204233](https://doi.org/10.1109/ICC.2003.1204233)
12. Riegel M, Tüxen M (2002) Mobile SCTP transport layer mobility management for the internet. In: Proceedings of the SoftCOM, pp 305–309
13. Snoeren AC, Balakrishnan H (2000) An end-to-end approach to host mobility. In: Proceedings of the 6th annual international conference on mobile computing and networking, pp 155–166. doi:[10.1145/345910.345938](https://doi.org/10.1145/345910.345938)
14. Fu S, Ma L, Atiquzzaman M et al (2005) Architecture and performance of SIGMA: a seamless mobility architecture for data networks. *IEEE Int Conf Communi (ICC 2005)* 5:3249–3253. doi:[10.1109/ICC.2005.1495024](https://doi.org/10.1109/ICC.2005.1495024)
15. IETF RFC2960 (2000) Stream Control Transmission Protocol
16. Stewart Randall R, Qiaobing Xie (2003) Stream control transmission protocol (SCTP): a reference guide. Tsinghua University Press, Beijing
17. Hong KH, Lee SK, Kim LY et al (2009) Cost-based vertical handover decision algorithm for WWAN/WLAN integrated networks. *EURASIP J Wireless Commun Netw* 2009:15. doi:[10.1155/2009/372185](https://doi.org/10.1155/2009/372185)
18. IETF draft-ietf-tsvwg-addip-sctp-08 (March 2004) Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration
19. IETF draft-sjkoh-msctp-01 (October 2005) Mobile SCTP (mSCTP) for IP Handover Support
20. IETF RFC3237 (2002) Requirements for Reliable Server Pooling

21. Koh SJ, Chang MJ, Lee M (2004) mSCTP for soft handover in transport layer. *IEEE Commun Lett* 8(3):189–191. doi:[10.1109/LCOMM.2004.823432](https://doi.org/10.1109/LCOMM.2004.823432)
22. Noonan J, Perry P, Murphy J (2004) Client controlled network selection. Fifth IEEE international conference on 3G mobile communication technologies, pp 543–547. doi:[10.1049/cp:20040734](https://doi.org/10.1049/cp:20040734)
23. Kashihara S, Nishiyama T, Iida K et al (2004) Path selection using active measurement in multi-homed wireless networks. In: *Proceedings of the 2004 international symposium on applications and internet*, pp 273–276. doi:[10.1109/SAINT.2004.1266127](https://doi.org/10.1109/SAINT.2004.1266127)
24. Funasaka J, Ishida K, Obata H et al (2005) A study on primary path switching strategy of SCTP. In: *Proceedings of autonomous decentralized systems 2005 (ISADS 2005)*, pp 536–541. doi:[10.1109/ISADS.2005.1452131](https://doi.org/10.1109/ISADS.2005.1452131)
25. Chebbine S, Chebbine M T, Obaid A et al (2005) Framework architecture and mathematical optimization of vertical handover decision on 4G networks using mSCTP. In: *IEEE international conference on wireless and mobile computing, networking and communications, 2005 (WiMob'2005)* 2:235–241. doi:[10.1109/WIMOB.2005.1512876](https://doi.org/10.1109/WIMOB.2005.1512876)
26. Chang M, Lee M, Koh S (2004) Transport layer mobility support utilizing link signal strength information. *IEICE Trans Commun* 87(9):2548–2556
27. Shi J, Jin Y, Guo W et al (2004) Performance evaluation of SCTP as a transport layer solution for wireless multi-access networks. *IEEE Wirel Commun Netw Conf 2004 (WCNC 2004)* 1:453–458. doi:[10.1109/WCNC.2004.1311587](https://doi.org/10.1109/WCNC.2004.1311587)
28. Ma L, Yu F, Leung VCM (2004) A new method to support UMTS/WLAN vertical handover using SCTP. *IEEE Wirel Commun* 2004:453–458
29. Ma L, Yu F, Leung VCM (2005) Modeling SCTP throughput in integrated WLAN/cellular networks. *IEEE* 2005:3445–3449
30. Song JK (2005) Performance evaluation of handoff between UMTS? 802.11b based on mobile IP and stream control transmission protocol. Master Thesis of North Carolina State University
31. IETF RFC2246 (1999) The TLS Protocol Version 1.0
32. IETF RFC3436 (2002) Transport Layer Security over Stream Control Transmission Protocol
33. IETF RFC3554 (2003) On the Use of Stream Control Transmission Protocol (SCTP) with IPsec
34. IETF draft-hohendorf-secure-sctp-15 (2012) Secure SCTP

Chapter 8

Mobility Management at Application Layer

The mobility management technology at the application layer is to provide location management, handover control, and security functions at the application layer. The advantages of this lie in hiding the variance of the heterogeneous access networks, without modifying the network infrastructures, but with greater flexibility and extensibility.

The Session Initiation Protocol (SIP) [1] and H.323 [2] define the mobility support capabilities at the application layer. SIP is chosen as the representative protocol in this chapter and is analyzed based on the reference model introduced in Chap. 2.

8.1 Overview

The application layer has inherent advantages in mobility support, including the following:

- Hiding the variance of the heterogeneous access networks. The application layer mobility management resides at the highest protocol layer and thus can provide mobility support that is independent of the access technologies.
- Providing better support for high-level mobility. As mentioned in Chap. 1, the high-level mobility objects encompass the various facets and targets of the user- and service-related mobility, which often depend on the service implementation at the application layer. Therefore, mobility support for the application layer can be integrated with the service provision for better support of the high-level mobility.

SIP and H.323 define mobility support capabilities at the application layer, although neither of them was originally designed for mobility. They extend the

location management, handover control, security, and interoperability functions for the base protocols. SIP was originally defined by the Internet Engineering Task Force (IETF) as a multimedia session control protocol, based on which some necessary extensions were proposed for mobility support capabilities. H.323 was standardized by the ITU-T and has no mobility support capability in its base protocol suite. H.501 [3], H.510 [4], and H.530 [5] were proposed to define the mobility requirements, mobility architecture, mobility management procedures, protocol messages, and security procedures for mobility in H.323.

The difference between SIP and H.323 in terms of their research backgrounds is that SIP focuses on the Internet field, while H.323 was designed for the evolution of traditional telephone networks. Their mobility support functions also show obvious differences because of their different design principles and network architectures, as summarized in Table 8.1.

Table 8.1 Comparison of mobility support functions of SIP and H.323

Mobility support function	SIP	H.323
Security mechanism	• Registrar is responsible for registration and authentication; • Mobile terminals send REGISTER message for registration and authentication; • Adopts the digest, stateless, and challenge-based authentication procedures defined in HTTP; • Adopts TLS for security transmission.	• Registration and authentication are implemented by gatekeeper (GK), home location function (HLF), visitor location function (VLF), and authentication function (AuF); • Hop-by-hop symmetric encryption is used to ensure security in the transmission of registration and authentication information; • Uses IPSec-based key distribution.
Location management	• Implemented through location registration; • Location server is responsible for storing and updating the location information; • Redirect server and proxy server use location service to locate the possible location of the callee; • No paging function in the base protocol; • SIP paging based on the integration of L2 paging and the application layer.	• Periodic location information updates are executed between the mobile terminal and the GK through registration, admission, and status (RAS); • Location update and management are implemented through the location information exchanged between the different administration domains, which are composed of GK, HLF, VLF, and AuF; • No paging function in the base protocol.
Handover control	• Supports handover through a re-INVITE message; • Only supports single-side handover based on UDP.	• For handover without change in network point of access (NPoA), mechanisms in the lower protocol layers are used to provide the handover function; • For handover involving a change of NPoA, H.323 does not define the necessary handover function.

In this chapter, SIP will be used as a typical protocol for mobility management at the application layer. The mobility control functions of SIP will be analyzed in detail based on the reference model defined in Chap. 2.

8.2 Typical Protocol: SIP

8.2.1 Overview

SIP is originally defined as an application layer control protocol for creating, modifying, and terminating multimedia sessions. The major network elements of SIP could be divided into two types: user agents (UAs) and intermediate servers.

A user agent is a logical entity that can act as both a user agent client (UAC) and user agent server (UAS). A UAC is a logical entity that creates a new request and sends it out. A UAS is a logical entity that generates a response to a SIP request, and this response may accept, reject, or redirect the request [1].

The SIP servers include registrars, redirect servers, proxy servers, and location servers. A registrar is a server that handles the location information. It accepts the location information in the REGISTER requests and places them into the location server. A redirect server accepts a SIP request and directs the client to contact an alternate set of URIs. A proxy server is an intermediary entity that acts as both a server and a client for the purpose of making requests on behalf of other clients. Its job is to ensure that a request is sent to another entity closer to the targeted user. A location server is used by a SIP redirect or proxy server to obtain information about a called party's possible location(s) [1].

A simple structure of the SIP network is shown in Fig. 8.1. Actual SIP network deployment may be more flexible and complex.

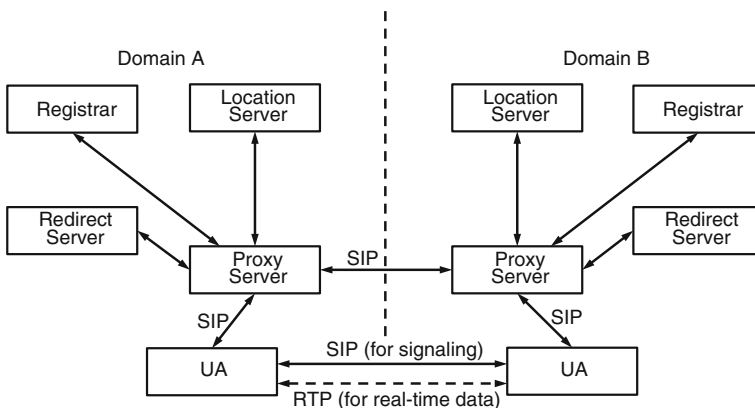


Fig. 8.1 Simple structure of a SIP network

More details of the SIP base protocol, the SIP network, and SIP messages can be found in RFC3261 [1].

As mentioned above, SIP is extended to provide mobility support capability. The mobility-related entities and functions include the following:

SIP provides a location management function. It defines the basic location service in its base protocol. Hierarchical deployment and distributed deployment are also introduced for better scalability. The location management processes of SIP will be introduced in Sect. 8.2.2.

SIP also provides a handover control function. It defines the basic mid-call handover support capability. Fast handover and soft handover mechanisms are also introduced for improved performance. The handover control processes of SIP will be introduced in Sect. 8.2.3.

SIP also provides some security mechanisms to support authentication, authorization, integrity, and confidentiality, and these mechanisms will be introduced in Sect. 8.2.4.

In SIP networks, the entities involved in mobility support include the mobile host (MH), the correspondent host (CH), and the SIP servers (including the SIP registrars, the location servers, and the proxy servers).

8.2.2 Location Management

(1) Basic location service in SIP

In SIP, the redirect server and the proxy server use the location information to learn of the current possible location of the destination user. The location information describes the binding of the address-of-record (AOR) uniform resource identifier (URI) and multiple contact addresses, where the AOR URI is often regarded as the public home address of the user, while a contact address represents the current available location of the user.

A registration operation is used in SIP to associate the AOR URI with the contact addresses. The registrar server writes and maintains this association, which is also called a binding, in a database (which is called the location service in SIP), where it can be used by the proxy server and the redirect server. The users send *REGISTER* messages to the registrar server to update these bindings.

Figure 8.2 shows the basic location service in SIP, including the location registration operation and the use of this operation in call setup. In steps a and b in Fig. 8.2, the MH uses a *REGISTER* message to perform location registration. Steps ①–⑤ in Fig. 8.2 show the call setup procedure, and this also illustrates the pre-call mobility support capability of SIP, where the MH acquires a new address prior to receiving or making a call.

The MH simply re-registers with its home registrar each time that it obtains a new address. When the CH subsequently makes a call to the MH through an *INVITE* message, the SIP server should provide the function of redirecting this

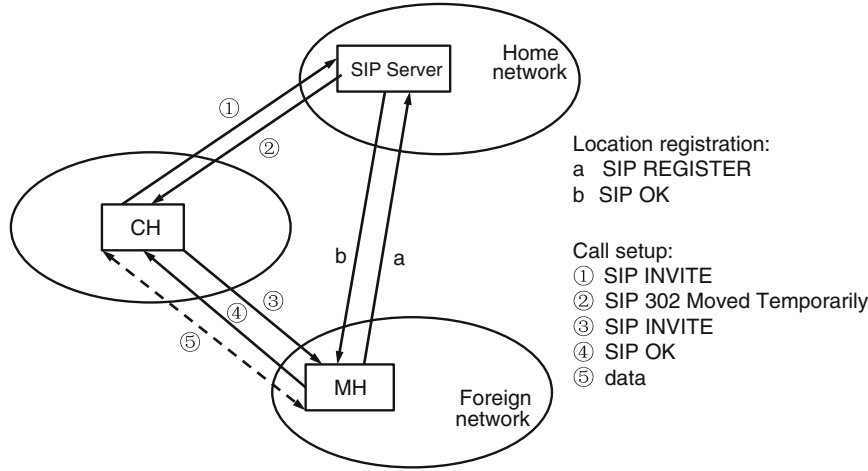


Fig. 8.2 Basic location service in SIP

INVITE message to the current location of the MH, which is implemented through the location service in SIP. In fact, the proxy server or the redirect server queries the registrar to acquire the location binding information.

Registration is just one way to implement location services in SIP. Other mechanisms have also been proposed for location management in SIP.

(2) Location management based on the integration of SIP and P2P

The scalable application layer mobility protocol (SAMP) is proposed in [6] for location management based on the integration of SIP and peer-to-peer (P2P) mechanisms. SAMP keeps track of the locations of the mobile nodes using the SIP mobility functions, where each MH performs a location registration operation and its location is tracked via SIP messages. However, the SIP servers in SAMP are organized as a P2P overlay network, which enables scalability, load balancing, and robustness features in mobility support.

In the SAMP architecture, the SIP servers form an overlay network by acting as peer nodes in the distributed hash table (DHT). SAMP uses this DHT for distributed location management functions. Each key in the DHT corresponds to the location information of one MH in SAMP, and the keys are maintained at the SIP servers. The SIP servers perform the functions of a registrar along with those of the proxy/redirect servers [6].

A. Location registration

Discovering the anchor SIP server is the first step for location registration. The anchor SIP server acts as an entry point to trigger a location registration procedure in the P2P overlay network. It also maintains the mapping information between the MN's SIP URI and the current CoA.

The MH discovers the anchor SIP server by broadcasting a *REGISTER* message carrying its home URI in the *From* field and the newly assigned IP address as its care-of address (CoA) in the *Contact* field. The anchor SIP server is randomly selected from the servers that provide answers for load balancing, along with considerations of delay constraints.

After the anchor SIP server is found, the MH will perform location registration with its home SIP server. The *REGISTRATION* message is then sent from the MH to its anchor SIP server and is then forwarded to its home SIP server via P2P application layer routing.

The location registration process in SAMP has a hierarchical manner, in which the anchor SIP server maintains the mapping information between the MN’s SIP URI and the current CoA, while the home SIP server maintains the binding information between the MN’s home SIP URI and the anchor SIP server’s address.

B. Session establishment

An explicit session establishment procedure is needed in SIP for session-based multimedia applications. This procedure is originated by the CH.

The CH sends an *INVITE* message to the overlay network through the CH’s anchor SIP server. This *INVITE* message is then routed to the MH’s home SIP server via P2P application layer routing. The MH’s home SIP server then forwards the message to the MH’s anchor SIP server, because it can find the current location of the MH using its registrar function. Because the MH’s anchor SIP server knows the CoA of the MH, it then relays the *INVITE* message to the MH.

Finally, the MN answers with a suitable response message, and the SIP session is established. After session establishment, the CN and the MN can communicate directly without involving the overlay network.

(3) SIP Paging

SIP paging is designed for an MH in the idle state. When a call is made with a particular MH as the destination, SIP paging is used to wake up that MH.

Several SIP paging mechanisms have been proposed in [7–9], as summarized in Table 8.2. Although different entities are involved for paging, a hierarchical

Table 8.2 SIP paging solutions

	Solution in [7]	Solution in [8]	Solution in [9]
Entities involved in SIP paging	SIP proxy servers	SIP proxy server with TA and PA, AR, AP	Paging location update server
Architecture	Hierarchical	Hierarchical	A paging location update server for each SIP domain
Paging method	At the application layer, using multicast <i>INVITE</i>	Integration of L2 and L3 paging	Integration of L2 and application layer paging

architecture is most often used as a solution. For example, [7] assumes that the SIP proxies are organized hierarchically, i.e., with a proxy deployed for each wireless network, region, cell cluster, and base station. In the paging architecture given in [8], an SIP proxy with a tracking agent (TA) and a paging agent (PA) handles the paging of the domain, which can then be partitioned further into multiple regions where each one is handled by an access router (AR) and can then be partitioned into multiple cells, with each one handled by an access point (AP).

Based on the hierarchical architecture, the paging operation will be started by the entity that has the most recent registration for the MH's location. If this entity does not receive the expected response within a certain timeout period, then the paging process will be restarted at the upstream paging entity.

With regard to the paging method, a multicast *INVITE* at the application layer, integration of the L2 and L3 paging, and integration of the L2 and application layer paging can all be used.

8.2.3 Handover Control

(1) Basic SIP-based handover control

SIP-based handover is also known as mid-call mobility support. The process is implemented through a re-*INVITE* message sent from MH to CH during an ongoing session.

Figure 8.3 shows the SIP-based handover procedure. The re-*INVITE* message sent from the MH to the CH uses the same call identifiers that were used in the original call setup. The newly assigned IP address (i.e., the new CoA) of the MH is carried in the *contact* field of the re-*INVITE*, telling the CH to send subsequent data

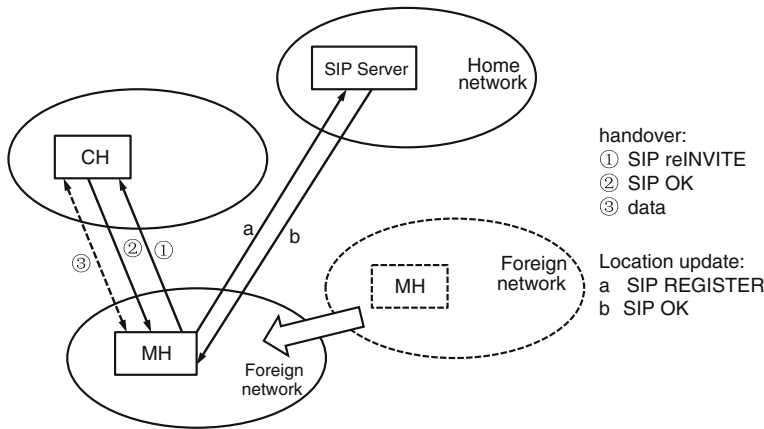


Fig. 8.3 SIP-based handover

packets to the new address. In addition, to redirect the data traffic flow, it also indicates the new address in the session description protocol (SDP) field, where it specifies the transport address.

Of course, the procedure described above is only used to maintain the continuity of an ongoing session. Registration to the home SIP server remains necessary to update the location of the MH, so that new calls can be forwarded to the MH. Figure 8.3 shows the location update procedure as indicated by steps *a* and *b*.

SIP-Based handover suffers from unsatisfactory performance for the following reasons:

- The detection of the mobility triggering events that happen in the lower layers (e.g., the link going down event) at the application layer incurs greater handover delays and thus more packet losses during handover.
- The IP address acquisition procedure and the authentication, authorization, and accounting (AAA) of the MH also incur handover delays and hence packet losses.

Therefore, performance improvements are required for SIP-based handover to ensure the application QoS.

(2) Performance improvement for SIP-based handover

Performance optimization of the SIP-based handover process can be addressed from two perspectives. Fast handover is used to reduce the handover delay, and soft handover follows the make-before-break principle and aims to reduce the packet losses during handover.

A. Fast handover based on SIP

Some solutions [10–13] have been proposed to reduce the handover delay in SIP-based handover. These solutions made their improvements from the procedures that introduce the different parts of the handover delay.

Miao et al. [10] focus on the handover delay caused by host mobility and especially on the delays that are accumulated during movement detection, duplicate address detection (DAD), SIP session re-establishment, and AAA procedures. The paper proposes a method that uses movement anticipation, tunneling, and AAA context transfer to alleviate the handover delays [10].

Mosawi et al. [11] combine 802.21 media independent handover (MIH) smart triggers with the SIP mechanism to provide a fast handover solution. The MIH functions are incorporated into the client architecture to interact with the SIP Intelligent Handover (SIPIHO) control process. The main purpose of the process is to reduce the handover latency that are introduced by scanning, association, and authentication and by updating the CH with the new IP address.

Cha et al. [12] use cross-layer interworking to reduce the handover delay. It proposes cross-layer fast handover for SIP (CF-SIP), which performs the movement detection and IP address reservation processes using the modified link layer

information in wireless networks, and proactively creates a new SIP session with the reserved IP address [12].

Dutta et al. [13] present several ways to achieve fast handover, which are designed to handle delayed SIP re-INVITE messages because of the distance between the CH and the MH and the congestion in the network. The possible solutions include the following: use of a SIP registrar and a real-time transport protocol (RTP) translator or network address translator (NAT), use of the outbound proxy, use of a multicast agent, and use of a back-to-back user agent (UA) or B2BUA as a mobility agent.

Table 8.3 compares the above fast handover solutions based on SIP.

B. Soft handover based on SIP

The soft handover provides a solution that can deal with the high handover delays and consequent packet losses. Some solutions for SIP-based soft handover were designed in [14–16]. Packet duplication and bicasting methods are often used to reduce the packet losses during handover, for which some extension headers have been designed for the SIP re-INVITE message used for handover. Table 8.4 summarizes these solutions.

The soft handover mechanism proposed in [14] is used here as an example of packet duplication-based solutions. The MH is assumed to be equipped with heterogeneous network interfaces and can receive and transmit packets through more than one of these interfaces simultaneously. The proposed soft handover

Table 8.3 Comparison of different solutions for fast handover based on SIP

Solution	Alleviated delay	Components involved	Protocol layers involved
[10]	The delay accumulated during the movement detection, DAD, and AAA procedures.	No additional components other than SIP entities	Application layer
[11]	The delays introduced by scanning, association, and authentication and by updating the CH with the new IP address.	MIH function is incorporated into the mobile terminal	MIH at L2.5 beside the application layer
[12]	The delays of movement detection and IP address allocation.	No additional components other than SIP entities	Modified link layer for application layer optimization
[13]	The delayed SIP re-INVITE message required because of the distance between the CH and the MH and the congestion associated with routing in the network.	The RTP translator deployed for each domain along with the SIP entities	Complete application layer solution

Table 8.4 Comparison of the solutions for soft handover based on SIP

Solution	Method	Extension to SIP
[14]	Packet duplication	JOIN header field in <i>INVITE</i> message
[15]	Packet duplication	JOIN header field in <i>INVITE</i> message
[16]	Bicasting	HANDOVER header field in <i>INVITE</i> message

procedure is initiated by the MH, but is executed at the base stations, which act as the gateways to the Internet in access networks.

Figure 8.4 shows the message sequence used in SIP-based soft handover defined in [14]. The call is established using MH_IF_I through BS_I, based on the conventional SIP call setup procedure. With the movement of the MH, another network interface (MH_IF_II in Fig. 8.4) is activated and is assigned a new IP address. An *INVITE* message with a JOIN header field is then sent from MH_IF_II to BS_I, carrying the new IP address in the contact field and the updated SDP parameters, and using the same call ID as used in the original session. Then, for a transient period, the MH sends and receives duplicate RTP packets through both MH_IF_I and MH_IF_II. The packet filters at the media gateway, and the MH discards the duplicate RTP packets. When the MH begins to receive the RTP packets through MH_IF_II, it sends a re-*INVITE* message to the CH to perform the handover. When a duplicate packet reaches the interface MH_IF_II, the connection from MH_IF_I is released by sending a *BYE* message to BS_I. Finally, the MH must register its new location information with the home SIP server.

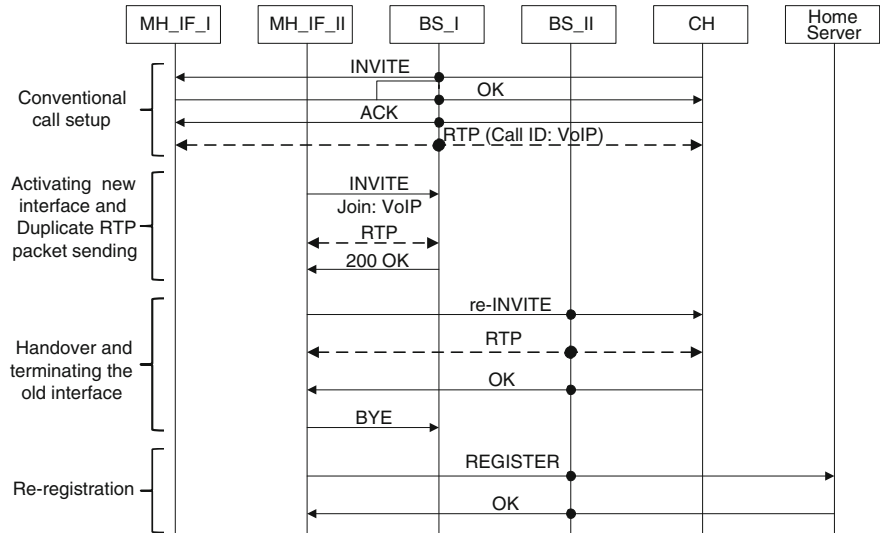


Fig. 8.4 Message sequence in packet duplication-based SIP soft handover

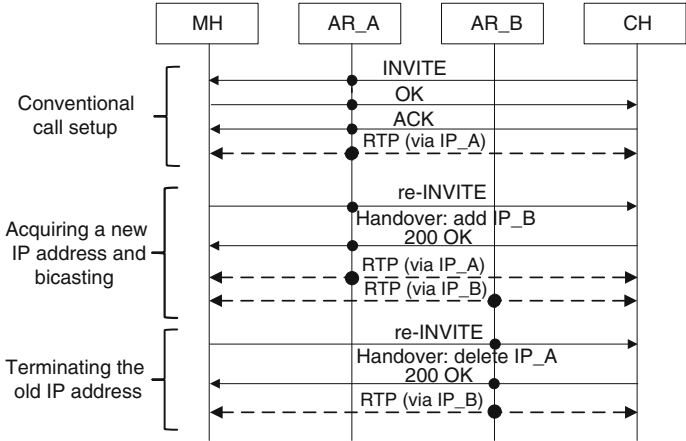


Fig. 8.5 Message sequence in bicasting-based SIP soft handover

The soft handover mechanism proposed in [16] is used as an example of soft handover based on bicasting of media streams. It defines an extension header **HANDOVER** field to be inserted into the re-INVITE message. This header instructs the CN to add or delete the IP address used for the SIP signaling and media streams, i.e., to start or stop bicasting.

Figure 8.5 illustrates the signaling procedure for soft handover proposed in [16]. The MH initially uses IP address A (IP_A) to establish the communication session. When it detects the availability of AR_B and is assigned a new address, IP_B, the MH will send an SIP re-INVITE message that contains the **HANDOVER** header field to indicate that IP_B is to be added. Then, the CH starts bicasting to the MH using IP_A through AR_A and IP_B through AR_B simultaneously. As the MH moves further into the coverage area of AR_B and detects the oncoming unavailability of AR_A, it will send another SIP re-INVITE message carrying the **HANDOVER** header field to stop the bicasting. After the corresponding 200 OK message is received, the MH and the CH use only IP_B for communication. Of course, registration with the home SIP server is also necessary to update the location information of the MH, but this is not included in Fig. 8.5.

8.2.4 Security Mechanism

SIP provides limited security mechanisms to guarantee authentication, confidentiality, and integrity.

(1) HTTP challenge-based authentication

SIP adopts a stateless, challenge-based authentication mechanism that is based on HTTP authentication. When a proxy server or a UA receives a request, it will

Table 8.5 Responses and header fields used in different authentication scenarios

Authentication scenario	Responses	Header fields
User–user authentication	401 status code: • Unauthorized response message; • Used by the UAS to reject the request and challenge the originator to provide credentials; • Can be used by the UAS, the registrar, and the redirect server.	WWW-Authenticate: • Included in the 401 (unauthorized) response message; • The field value consists of at least one authentication challenge.
		Authorization: • Included in the reorganized request message; • The field value contains the authentication credentials of a UA.
Proxy–user authentication	407 status code: • Proxy Authentication Required response message; • Used by the proxy server to reject the request and challenge the originator to provide credentials.	Proxy-Authenticate: • Included in the 407 (Proxy Authentication Required) response message; • The field value contains an authentication challenge.
		Proxy-Authorization: • Included in the reorganized request message, allowing the client to identify itself (or its user) to a proxy that requires authentication; • The field value consists of credentials that contain the authentication information of the UA for the proxy and/or the realm of the resource being requested.

challenge the initiator of the request to provide assurance of the initiator’s identity. When the originator has been identified, the recipient of the request will then further ascertain whether or not this user is authorized to make the request in question [1].

The HTTP authentication process can be used for user–user authentication and proxy–user authentication. User–user authentication is used by the UAS when it receives a request from a UAC to authenticate the originator before the request is processed. Proxy–user authentication is used by a proxy server to authenticate the originator when it receives a request from a UAC.

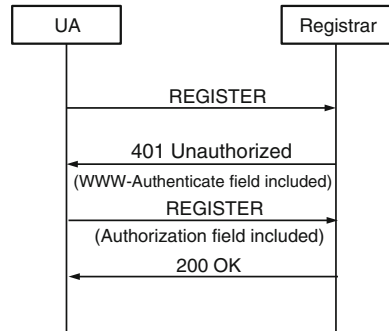
Some response status codes and necessary header fields that carry challenges and credentials are defined for the above authentication scenarios. Table 8.5 summarizes these header fields and responses.

Figure 8.6 shows an example of the HTTP challenge-based authentication procedure between a UA and a registrar.

(2) Message integrity and confidentiality

Message integrity and confidentiality in SIP can be guaranteed by using of some lower layer schemes or the Secure Multipurpose Internet Mail Extensions (S/MIME) specification.

Fig. 8.6 An example of HTTP challenge-based authentication between a UA and a registrar



For lower layer security schemes, SIP can use transport or network layer-based security to encrypt the signaling traffic, guaranteeing both message confidentiality and integrity. Two popular options for providing this security are transport layer security (TLS) at the transport layer and IPSec at the network layer, which are often used on a hop-by-hop basis.

For S/MIME, the rationale exists in that the SIP messages follow the MIME format and S/MIME is simply the specification used to provide security for MIME. S/MIME can provide end-to-end confidentiality and integrity for SIP message bodies, along with mutual authentication.

8.3 High-Level Mobility Support of SIP

Besides the above mobility support functions for terminal mobility, SIP has more mobility support capabilities for high-level mobility objects than any of the other mobility management protocols.

8.3.1 Personal Mobility Support of SIP

As described in Chap. 1.2.2, personal mobility is a type of mobility in scenarios where the user changes the terminal that is used for network access at different locations.

In personal mobility based on SIP, two important capabilities are provided. One is the ability to address a single user located at different terminals using the same logical address, while the other is the ability to allow the users to maintain their media session even while they are changing terminals [7].

The ability to address a single user located at different terminals using the same logical address is used for users that are reachable via multiple terminals, such as a traditional public switched telephone network (PSTN) telephone, a mobile phone, a

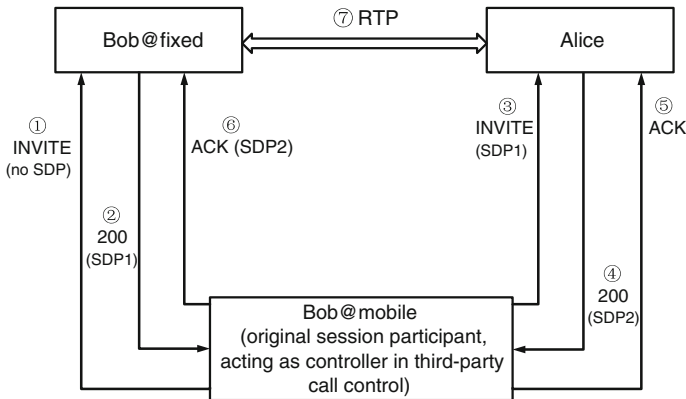


Fig. 8.7 Third-party call control for personal mobility

video terminal, and a PC. The mapping between the logical address and the possible terminals is important for supporting this capability. In this case, both the 1-to-n (one address, many potential terminals) and m-to-1 (many addresses reaching one terminal) mappings are useful [7]. SIP forking proxies are used to allow the user to be reached at any of the potential devices using the same name.

The ability to maintain a media session while the user is changing terminals is used when the user wants to move a session or part of a session from one device to another. SIP defines two extensions as solutions to support this capability: third-party call control [17] and the *REFER* mechanism [18]. We use an example to introduce these two methods. In our example, we assume that Alice is in a session with Bob at a mobile host (Bob@mobile) and wants to move the session to Bob at a fixed host (Bob@fixed).

Figure 8.7 illustrates the session mobility support based on third-party call control mechanism [17]. The original session participant (Bob@mobile, which also acts as the controller in third-party call control) sends an *INVITE* request to the new session destination (Bob@fixed), which indicates the session parameters, such as the IP address of the remote session participant (Alice). Bob@mobile also sends Alice the session description that is generated by Bob@fixed, so that Alice then sends media streams to Bob@fixed instead of Bob@mobile. Bob could also split the session into components, with different receivers for each media type. Obviously, the disadvantage of this solution is that the original session participant must remain involved in the session, as they will be contacted to change or terminate the session [7].

Figure 8.8 shows the solution based on the *REFER* method [18], which explicitly transfers the session to the new destination. In this case, Bob@mobile simply sends a *REFER* request to Alice, indicating that she should contact Bob@fixed. Alice then negotiates a session with Bob@fixed using the regular *INVITE* exchange. Alternatively, Bob can also send a *REFER* request to Bob@fixed, asking him to invite Alice [7].

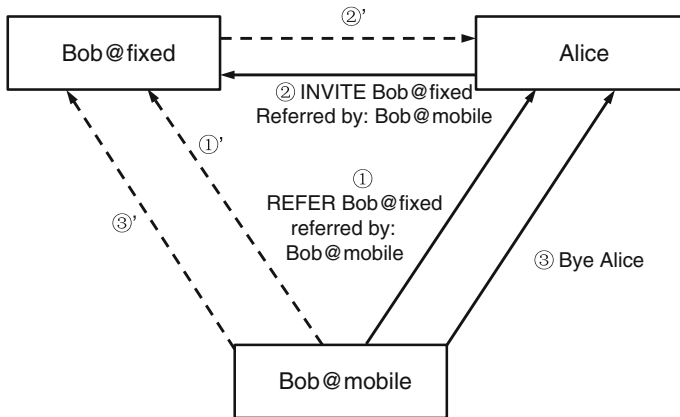


Fig. 8.8 REFER mechanism for personal mobility

8.3.2 Service Mobility Support of SIP

SIP provides service mobility that allows users to maintain access to their services even when moving or changing their devices or network service providers [7]. Two important issues must be considered here.

The first issue is the ability to ensure that users have access to all their subscribed services, regardless of their point of attachment to the network. This issue can be realized based on service profile portability. The process can be supported either by having the user carry his service profile information with him, or by having network storage in which the profile information is stored in specified servers with downloading, updating, and synchronization methods.

The second issue is the ability to maintain an adequate QoS for the duration of a session, regardless of any network changes. One possible SIP approach to QoS support is through appropriate resource allocation (using the *COMET* method) during session initiation and handovers [19].

8.4 Applications of Mobility Support of SIP

8.4.1 SIP-Based Mobility Support in Vehicular Networks

SIP is considered to be one possible solution for mobility support in vehicular networks with high mobility features.

For example, SIP is used in vehicular ad hoc networks (VANETs) to support secure and QoS-aware handover for voice over IP (VoIP) communications [20].

The network architecture is assumed to use SIP components to allow access to the VoIP applications. An onboard unit (OBU) is deployed on the vehicle, acting as

the SIP UA and as the endpoint of a call. A roadside unit (RSU) is used as the OBU attachment point to the network. The VANET is composed of several IP zones that are deployed with SIP servers, such as the registrar server, the location server, the proxy server, the redirect server, and the dynamic host configuration protocol (DHCP) server.

VoIP communication is performed between vehicles that belong to different zones. When a vehicle moves across different RSUs that belong to the same zone, a MAC layer handover is performed by the RSU. When the vehicle moves from one zone to another, SIP-based handover at the application layer should also be executed, basically by following the handover procedure introduced in Sect. 8.2.3.

For security protection, Triki et al. [20] use an established secure SIP session to exchange new session keys during the pre-registration phase, which provides a mutual authentication mechanism and protects against known SIP attacks. With regard to the QoS issue, Triki et al. [20] use a set of parameters (e.g., the signal strength, RSU memory usage, the road map, and the vehicle speed and direction) to ensure that the voice quality is maintained during handovers, and the number of resources consumed in each SIP session by handover execution is reduced.

8.4.2 SIP-Based Mobility Support in Satellite Communication Networks

SIP is also considered as a possible technology to provide mobility support in satellite communication networks. For example, in the satellite-based communications systems within IPv6 networks (SATSIX) project, SIP has been chosen as one of the possible protocols to be used to provide integration of both mobility and QoS at the application layer.

SATSIX is a R&D project funded by the European Union, with the aim of enabling efficient delivery of broadband services using satellite links [21]. The SATSIX project introduces a mobility architecture, which defines the joint deployment of SIP and MIPv6 for some QoS enabled applications. The mobility part of the architecture will be managed by MIPv6 and its optimizations (hierarchical MIP (HMIP), fast MIP (FMIP), and F-HMIP) for the TCP connections and by SIP-based mobility [22] for real-time communications over UDP [23].

Assuming the scenario of an MH with an ongoing SIP session (e.g., VoIP) and some TCP connections (ftp, Web browsing) with the same CH moving from its home network to a visited network, the mobility support functions for these different applications are handled by SIP and MIPv6, respectively.

For location management, when the MH moves into the visited network, it will register its new location with both the HA in MIPv6 and the home SIP server. Indeed, some solutions are required to colocate the HA and the SIP server in the same device or to allow the home SIP server to query the HA [23].

For the handover operation, the MH will simultaneously send a BU message to the HA and a re-*INVITE* message to the CH to handle the handover. After the BA and the SIP OK are received, the TCP connections can continue via the HA or directly with the CN if the route optimization is activated and the SIP session can continue directly between the MN and the CN [23].

A SIP-based soft handover mechanism is proposed in [24] for satellite-terrestrial integrated networks. It is used for MHs that can connect to a mobile satellite system (MSS) and the universal mobile telecommunications system (UMTS) simultaneously. The MHs are assumed to use the UMTS for higher data rates and lower costs, but handover to the MSS when moving into areas without UMTS coverage (e.g., deserts and oceans).

The satellite access gateway (SAG) in the satellite station and the terrestrial access gateway (TAG) in the GPRS network are added to the integrated network architecture as SIP proxies of the MH for SIP soft handover. They are responsible for assisting the MH in performing session control, registration, service access, and seamless handover procedures.

The soft handover procedure in [24] also uses duplicated data streaming during handover to reduce the packet losses, as discussed in Sect. 8.2.3. The data duplication is performed at the SAG (for handovers from MSS to UMTS) or the TAG (for handovers from UMTS to MSS).

References

1. IETF RFC 3261 (2002) SIP: session Initiation Protocol
2. ITU-T Recommendation H.323 (2003) Packet-based multimedia communications systems
3. ITU-T H.501 (2002) Protocol for mobility management and intra/inter-domain communication in multimedia systems
4. ITU-T H.510 (2002) Mobility for H.323 multimedia systems and services
5. ITU-T H.530 (2002) Symmetric security procedures for H.323 mobility in H.510
6. Pack S, Park K, Kwon T et al (2006) SAMP: scalable application-layer mobility protocol. *IEEE Commun Mag* 44(6):86–92. doi:[10.1109/MCOM.2006.1668425](https://doi.org/10.1109/MCOM.2006.1668425)
7. Schulzrinne H, Wedlund E (2000) Application-layer mobility using SIP. *Mobile Computing and Communications Review* 4(3):47–57
8. Sarikaya B, Ozugur T (2005) SIP paging of wireless LAN hosts for VoIP. In: *Proceedings of IEEE Vehicular Technology Conference*, 4. Spring, 2345–2348. doi:[10.1109/VETECS.2005.1543754](https://doi.org/10.1109/VETECS.2005.1543754)
9. Watanabe F, Wu G (2002) Method and apparatus for an SIP based paging scheme. US Patent 7020440
10. Miao W, Yu-Jun Z, Jun L (2006) A Fast Handover Solution for SIP-based Mobility. In: *International Conference on IEEE Wireless and Mobile Computing, Networking and Communications (WiMob'2006)*, 315–320. Doi:[10.1109/WIMOB.2006.1696341](https://doi.org/10.1109/WIMOB.2006.1696341)
11. Al Mosawi T, Shuaib H, Aghvami A H (2009) A Fast Handover Scheme Based on Smart Triggers and SIP. In: *IEEE Vehicular Technology Conference Fall (VTC 2009-Fall)*, 1–5. Doi:[10.1109/VETECF.2009.5378939](https://doi.org/10.1109/VETECF.2009.5378939)
12. Cha E, Lee K, Kim M (2007) Cross layer fast handoff for SIP. In: *IEEE 21st International Conference on Advanced Information Networking and Applications AINA'07*, 443–450. Doi:[10.1109/AINA.2007.48](https://doi.org/10.1109/AINA.2007.48)

13. Dutta A, Madhani S, Chen W, et al (2004) Fast-handoff schemes for application layer mobility management. In: 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC 2004 3, 1527–1532. Doi:[10.1109/PIMRC.2004.1368255](https://doi.org/10.1109/PIMRC.2004.1368255)
14. Banerjee N, Acharya A, Das SK (2006) Seamless SIP-based mobility for multimedia applications. *IEEE Network* 20(2):6–13. doi:[10.1109/MNET.2006.1607890](https://doi.org/10.1109/MNET.2006.1607890)
15. Zhang W, Wang J (2009) A vertical soft handoff scheme based on SIP in the ubiquitous wireless network. In: IEEE Fourth International Conference on Computer Sciences and Convergence Information Technology, ICCIT'09, 794–799. Doi:[10.1109/ICCIT.2009.55](https://doi.org/10.1109/ICCIT.2009.55)
16. Koh SJ, Hyun W (2008) mSIP: Extension of sip for soft handover with multicasting. *IEEE Commun Lett* 12(7):532–534. doi:[10.1109/LCOMM.2008.080335](https://doi.org/10.1109/LCOMM.2008.080335)
17. IETF RFC3725 (2004) Best Current Practices for Third Party Call Control (3pcc) in the Session Initiation Protocol (SIP)
18. IETF RFC3515 (2003) The Session Initiation Protocol (SIP) Refer Method
19. Popescu L (2003) Supporting multimedia session mobility using SIP. *Proc First Ann Commun Netw Ser Res Conf (CNSR)* 2003:122–123
20. Triki B, Rekhis S, Boudriga N (2011) Secure and QoS-aware SIP handover for VoIP communication in vehicular adhoc networks. In: IEEE 2011 7th International Wireless Communications and Mobile Computing Conference (IWCMC), 695–700. Doi:[10.1109/IWCMC.2011.5982631](https://doi.org/10.1109/IWCMC.2011.5982631)
21. Arnal F, Gayraud T, Baudoin C et al (2008) Ip mobility and its impact on satellite networking. *IEEE Adv Satell Mobile Syst, ASMS* 2008:94–99. doi:[10.1109/ASMS.2008.24](https://doi.org/10.1109/ASMS.2008.24)
22. Wedlund E, Schulzrinne H (1999) Mobility support using SIP. In: Proceedings of the 2nd ACM international workshop on Wireless mobile multimedia. *ACM*, 76–82. Doi:[10.1145/313256.313281](https://doi.org/10.1145/313256.313281)
23. Arnal F, Baudoin C, Jacquemin B et al (2008) Emulation of the SatSix mobility architecture. *IEEE Int Workshop on Satell Space Commun, IWSSC* 2008:198–202. doi:[10.1109/IWSSC.2008.4656785](https://doi.org/10.1109/IWSSC.2008.4656785)
24. Lan H, Hu X, Zhang X, et al (2012) A SIP-based Soft Handoff Scheme for Integrated Satellite-Terrestrial Networks. In: IEEE 2012 8th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM), 1–4. Doi:[10.1109/WiCOM.2012.6478371](https://doi.org/10.1109/WiCOM.2012.6478371)

Part III
Future Requirements, Challenges
and Directions

Chapter 9

Requirements, Challenges, and Trends

From its origins in cellular communication systems, mobility management has experienced the network evolution of the Internet and the mobile Internet and continues toward the future networks. It was treated as one of the most important technologies in specific networks (e.g., cellular systems, wireless mesh networks, 802.11 wireless networks) and continues to evolve as a fundamental capability that is essential to the network.

Increasing numbers of mobility scenarios are bringing about new requirements and thus new challenges for mobility management. This chapter firstly analyzes the scenario-specific requirements, from which the methodology-related challenges are derived, and then, the solution trends are introduced accordingly.

9.1 Scenario-Specific Requirements

The new technical requirements are considered from the application, terminal, and network perspectives, motivated by the observations that current networks show obvious application-driven features in which increasingly abundant applications bring about new changes in terminal technologies, which subsequently incur changes in the networks.

9.1.1 *The Application Perspective*

(1) Diversified mobile service

Short text messages and voice calls, which are offered by mobile communication systems, are considered as traditional mobile services. Thanks to ubiquitous mobile

data access and the popularity of powerful mobile devices, mobile services have become increasingly attractive and diversified. Today, almost all applications that previously ran on the PC platform have their counterparts on mobile device platforms, such as mobile searching, mobile payments, mobile browsing, mobile gaming, mobile reading, and mobile TV. Conversely, some native mobile applications (i.e., applications that launched initially on mobile device platforms) also have their counterparts on PC platforms.

The most important differentiating factor of mobile applications is user environmental awareness, e.g., location awareness or proximity awareness. This means that location information and/or environmental information can be used as an input to provide new service features and user experiences, such as location-based applications, SoLoMo (created by John Doerr to signify social, local, and mobile) applications, and mobile e-health services. Another example is by using real-time traffic information, a navigation application could find a route that can avoid a traffic jam and could then adjust the route dynamically according to the information received. Through knowledge of people in close proximity with similar interests, SoLoMo applications can enrich their users' social lives by making it easier to make new friends.

One most important enabler of such diversified mobile service is the powerful computing ability of mobile devices. For example, many handheld mobile devices have become multi-core, multi-RAT (radio access technology) devices equipped with storage capacities, and their operating systems are multitasking-enabled. This means that concurrent applications and services can be supported well on these mobile platforms. The mobile terminal not only works as a consumer of mobile services, but can also work as a provider of mobile services, e.g., for digital content creation and real-time content sharing. In contrast, cloud computing and HTML5 have also reduced the computing requirements on terminals by moving more computing-intensive tasks from the terminals to the network. All the above technological advances have boosted mobile service diversification.

From the application and service perspective, future mobility management schemes must consider the following new requirements, including mobile content sharing, and sharing of mobility-related context among different applications, e.g., identifiers, location information, and application contexts; performance improvement requirements under the application layers, e.g., at the network and transport layers (i.e., multiple interfaces (MIF) and Multipath Transport Control Protocol (MP-TCP)); and application security requirements, e.g., to protect not only the privacy but also the financial asset of the mobile service user.

(2) Mobile social network

A mobile social network (MSN) can be considered as a system that provides a variety of data delivery services involving the social relationships among the mobile users [1].

MSNs can be classified into two types: Web-based MSNs and decentralized MSNs [1].

- A Web-based MSN uses social network services (e.g., Facebook, MySpace, and microblogging Web sites) or mobile portal Web sites (e.g., 3G.net.cn) for acquiring information through mobile devices [2]. The mobile users communicate with the Web-based applications through the Internet, given the availability of wireless connectivity.
- In a decentralized MSN, a group of mobile users is formed to disseminate data among the group members without connecting to a centralized server. The mobile users communicate when they come into opportunistic contact using wireless technologies such as Bluetooth or WiFi [1].

MSN does not only mean access to social network services through mobile terminals. The introduction of mobile terminals into social networks brings further differences, which are summarized in Table 9.1.

Mobility, which is an important feature introduced by the mobile terminals, provides another method for social relationship formation in MSNs and thus can be used for efficient and effective data delivery services [1]. Location awareness is another prominent feature of MSNs because the location information of the mobile users can be provided conveniently by base stations (BSs) in cellular systems or through the Global Positioning System (GPS). This location information, when combined with the social network, can provide the MSN users with a variety of contextual services [1]. Therefore, as the popular term “SoLoMo” indicates, the MSN is a combination of a social network, a location-based service (LBS), and the mobile Internet.

Accordingly, the mobility-related requirements incurred by MSNs focus on identifying mobility patterns and applying them to protocol design.

Table 9.1 Differences between MSNs and classical social networks [1, 3]

	Classical social network	Mobile social network
Terminals	Often PCs	Often mobile terminals such as smartphones
Mobility	No	Yes
Location awareness	Only assistant information in applications	Crucial factor in applications
Social relationship definition	Defined by the users	Defined by the users, or analyzed based on mobility patterns
Data delivery mode	Content provider to user	Content provider to user in Web-based MSN, or user to user in decentralized MSN
Service display manner	Often on PC monitors through Web pages in browsers	Often on mobile terminals with small screens
Critical technology	Web technology	Web technology, together with terminal technologies (e.g., mobile operating systems, mobile widgets)

Mobility is an important feature in MSNs when compared with traditional social networks, which makes the mobility patterns an important factor for MSNs. The social relationships can be analyzed based on mobility patterns other than those defined by the users. The community in a decentralized MSN can be formed according to common interests and similar mobility patterns, where the mobility makes community detection in the MSN more complicated. Analysis of the mobility patterns provides a basis for mobility prediction, which can then be used to design efficient routing and data dissemination protocols.

Exploration of the mobility patterns raises another interesting issue, i.e., investigation of the mapping between the social network metrics (e.g., betweenness, closeness, and clustering coefficients) and the mobility parameters (e.g., contact duration and the amount of data to be transferred between encounters) and of the mapping between the virtual connection and the physical connection, may provide efficient methods for MSNs.

In addition, the mobility patterns in MSNs should take real human mobility patterns and behavior into account, because these mobile devices are carried by people.

9.1.2 *The Terminal Perspective*

(1) Group mobility

Group mobility is a mobility scenario in which a group of mobile nodes (MN) move together with varying topologies and attachment points while maintaining communication continuity. A group of terminals moving together and enjoying the wireless connectivity provided by a mobile relay is an example of group mobility. The group mobility scenario that is introduced by terminal cooperation can be considered as another example. Adjacent mobile terminals belonging to different users form a group and move together. The service provision could then be conducted based on their common interests and adjacent locations. For example, when the users on a bus request the same online video file, the different clips with their different coding formats (e.g., based on scalable video coding (SVC) or multi-description coding (MDC)) can be delivered to the different terminals, after which these clips can be shared between adjacent users in an ad hoc manner for better quality. The travel of the bus means that these terminals move together as a group. The topology inside the group changes dynamically as the users join and exit the group.

Group mobility is more complex than the traditional network mobility (NEMO) scenario. The similarity between group mobility and NEMO exists in a group of MNs moving together. However, the differences exist in the formation of the group, the dynamic features of the gateway node, the paths between the MNs and the gateway, and the attachment point to the external network, as shown in Table 9.2.

Table 9.2 Differences between group mobility and traditional NEMO

	NEMO	Group mobility
Formation of the group	The networking topology	The networking topology and the new service provision manner
Gateway	Remains unchanged as MR (mobile router)	Changes dynamically
Path composition between MNs and the gateway	One-hop wireless link in ordinary NEMO, or multi-hop in nested NEMO	Multi-hop ad hoc paths
Path diversity between MNs and MR	No	Yes
The attachment point to the external network	Changes because of mobility of MRs May be heterogeneous when MR is configured with multiple interfaces	Changes because of mobility of the gateway or changes in the gateway

Group mobility presents the following remarkable features:

- The groups may be formed explicitly or implicitly. In explicit groups, the MNs move together in an obvious manner, e.g., the terminals in a bus, train, or airplane. Implicit groups can be formed because of the service provision manner, and the similarity and dependency of the mobility metrics, e.g., the groups in MSN services, or the groups that form naturally in vehicular networks.
- Similarity in mobility metrics. This is an important feature of group mobility and also provides a method for the organization and maintenance of the groups.
- High dynamics in the group topology. This feature is related to other variable factors in group mobility, such as the attachment point of the group to the external network, the gateway nodes of the group, and the routing inside the group, which bring further challenges in mobility management for group mobility.

The technical requirements for group mobility therefore include the following:

- Modeling of group mobility behavior, which explores the group mobility metrics and patterns, for both explicit groups and implicit groups, considering both the physical network topology and the logical service provision manner and defining mobility metrics that range from traditional physical movement metrics to temporal and spatial mobility dependence metrics and then to more abstract group metrics.
- Dynamic organization of the groups, which aims to form groups based on the physical network topology, the virtual application topology, or the integration of the physical topology and the virtual topology.
- Mobility management based on dynamic groups, which handles the mobility management in granularity of the groups for efficient resource utilization and green communications, and handles any internal topology changes at the same time.

(2) Device-to-device communication

One device communicating with another device (or devices) directly via the air is the most straightforward and efficient method for device-to-device communication.

The walkie-talkie (produced by Motorola) was one of the earliest and most typical direct radio communication devices, which used an unlicensed spectrum and enabled half-duplex voice services without any infrastructure. Terminals from the TETRA (Terrestrial Trunked Radio) system, which uses a licensed spectrum, can still communicate with each other even when they are outside the network coverage area. For data communications, the IEEE 802.11 standard supports direct device communications in ad hoc mode and WiFi direct mode; both operate in the unlicensed spectrum. In the former, an ad hoc/mesh network is formed dynamically because of the terminal’s mobility. In the latter, one control device is selected before communication begins, and this device controls the subsequent communications. This is a new trend, which can be used for vehicle and roadside device communications in intelligent transport systems and for communication between consumer electronic devices, such as photograph sharing between digital cameras, or video playback to a TV from a smartphone. The progress made in direct communications to date is summarized in Table 9.3.

Based on their awareness of their proximity provided with the help of GPS, SoLoMo applications can exchange application data with each other via the mobile core network, but not directly between the terminals. This results in inefficient network resource usage, and the intensive GPS usage drains the terminal’s battery rapidly. With the aim of avoiding these problems, along with the network operators’ financial requirements, and driven by regulations from the USA and Europe for safety in public usage, working items related to direct communications between devices have been approved by the 3rd Generation Partnership Project (3GPP) within the Release 12 timeframe. The possible scenarios are portrayed in simplified form in Fig. 9.1 [4].

The left side depicts the in-coverage scenarios. The proximity awareness is achieved using LTE (Long-Term Evolution) radio, either directly or with the help of the network; direct communication between the devices (e.g., a and b) uses either

Table 9.3 Progress in direct communications

Examples of direct communications	Spectrum	Market/main purpose	Services offered	Communication method
Walkie-talkie like	Unlicensed	Commercial	Voice	Broadcast
Direct mode operation (DMO) in TETRA	Licensed	Public safety	Voice	Broadcast
Ad hoc/mesh	Unlicensed, or dedicated	Military	Data	Unicast
WiFi direct	Unlicensed	Consumer	Data, video	Unicast
Bluetooth-based	Unlicensed	Consumer electronics	Voice, data	Unicast

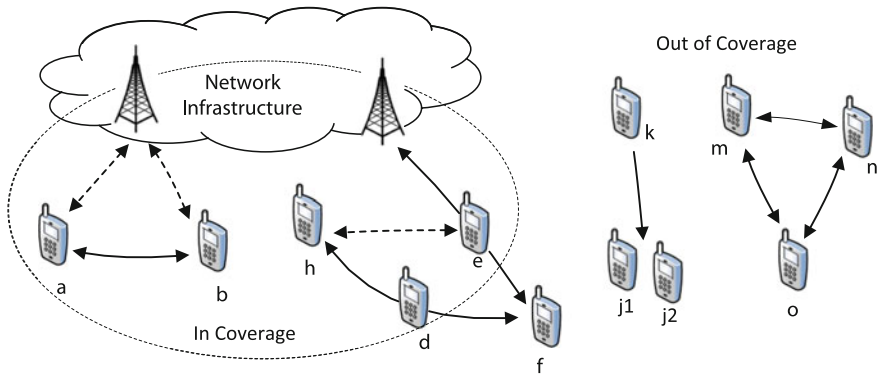


Fig. 9.1 Possible scenarios for device-to-device communications

the radio resources from the licensed spectrum or those from the unlicensed spectrum (e.g., WiFi). If a device (e.g., *f*) is in out-of-coverage area, but is in proximity to devices that are in coverage (e.g., *d* and *e*), then a device-to-device relay (via *d*) or a device-to-network relay (via *e*) can be used for *f* to communicate with the other devices or via the network, respectively. The out-of-coverage scenarios are depicted on the right side. One-to-many communication is possible; this means that one device (e.g., *k*) sends common information to many devices (e.g., *j1* and *j2*) at the same time.

The communications-related requirements can be categorized into proximity discovery and direct communication path establishment. From the mobility management perspective, technologies are needed to meet the specific requirements of more complex scenarios such as terminal relay to terminal, terminal relay to network, and out-of-coverage scenarios. Service continuity between the direct communication mode and the through-network mode is also an important requirement, which is used to ensure the quality of the user experience. The satisfaction of all these requirements offers the prospect of great challenges in future mobility management research.

9.1.3 The Network Perspective

(1) Mobility in a converged network

Initially, almost every single network was built independently to meet its own purpose, e.g., the public switched telephone network (PSTN) for fixed line voice communications, the public land mobile network (PLMN) for mobile voice communications, and the wireless local area network (WLAN) for wireless data access. However, more individual networks mean greater expenditure for the network operator, and paying the bills that come from each network also places a burden on customers.

To reduce expenditure and enhance the customer experience and customer loyalty, almost all operators are more eager than ever for network convergence. From a mobile network-centric perspective, network convergence can be classified into the following four scenarios, which will be briefly described below.

Different generation mobile network convergence refers to coexistence and interworking between different generations of mobile networks, because of the gradual deployment of new-generation networks and the gradual phaseout of the old networks. This would require not only voice service and data service continuity when the users move among the different generations of networks but also converged network architectures and technologies to achieve that continuity.

Mobile and wireless network integration is one of the most important trends in convergence. Existing work related to interworking between WLANs and mobile networks has been discussed in Sect. 5.7, including the intelligent WLAN (I-WLAN) architecture, the LTE network-centric integrated architecture, IP flow mobility, and non-seamless WLAN off-loading. The most recent issues have been studied by 3GPP, including performing WLAN network selection based on the policies defined by the WLAN operator (based on the HotSpot2.0 specification) and those defined by the mobile operator (based on ANDSF MO (access network discovery and selection function management object)); the network-based mobility mechanism with regard to the S2a reference point; and achieving efficient traffic off-loading to a WLAN when the terminal moves between different 3GPP radio networks.

Fixed and mobile network convergence has been studied for many years. The main goal in the beginning was to define a common network that achieved common subscription and billing management, common authentication and authorization, and common IP multimedia subsystem (IMS) service layers. Because of the introduction of femtocells and the popularity of WiFi routers, fixed and mobile network convergence occurs at the network edge. Terminal movements among the femtocell, the macro-cell, and the WiFi access network, not only to off-load traffic from the mobile core network to the fixed network, but also to achieve a common policy, QoS, and charging control architecture, present major tasks in the convergence process.

Satellite network interworking with mobile networks first happened for voice service interworking. Later, satellite links were used to backhaul the 3G radio access networks deployed on airplanes to the land-based core network. Nowadays, WiFi access to the Internet is already available in some airplanes. The affordable cost of satellite links means that interworking between mobile networks, wireless networks, and satellite links will become popular and thus will present more challenges.

Based on the above scenarios, more network requirements and technologies are emerging for mobility management. Referring to the protocol reference model given in Chap. 2, these requirements can be classified into control plane and user plane requirements. In the control plane, interoperation for QoS, new location management schemes such as distributed mobility management, security aspects related to autonomous authentication when accessing WLANs, and handover and

off-loading control technologies for simultaneous access to different types of radio are among the topics that need further study. In the user plane, the IP flow mobility and off-loading technologies also require further study.

(2) Mobility in heterogeneous and multilayered radio networks

The heterogeneous radio network (i.e., the HetNet [5], in the 3GPP context) refers to a radio network environment that consists of different types of radio cells, such as macro-cells, and low-power nodes (LPNs), e.g., microcells, picocells, femtocells, and relay nodes (please refer to Fig. 9.2) and that also consists of different radio access technologies, e.g., 3G, LTE. The major purpose of deploying a HetNet is to move the cellular network closer to the terminal and thus increase the spectrum efficiency and spatial use. Mobility management is one of the specific challenges faced by the HetNet. The deployment of multiple LPNs may increase the mobility management complexity, because the movement of a terminal may trigger frequent handovers when the terminal moves across the small coverage areas of the LPNs [5].

Recently, to increase data throughput and spectrum usage efficiency while dramatically reducing base station power consumption at the same time, some new radio network architectures have been proposed, such as the HCN (hypercellular network) [6], a novel architecture for LTE-B [7], and small cells [8]. The basic idea behind these architectures is C-plane and U-plane decoupling over the air interface, i.e., the control plane channel and the user plane channel use different spectrum bands and reside on different cells. One architecture option is to dedicate the 700 MHz band to cells that carry control plane messages only and allocate the 2.3 GHz band to cells that carry user plane data (refer to Fig. 9.2, where BS a and b operate the control plane cell and the user plane cell, respectively). By using C- and U-plane decoupling, the cells used for the U-plane only can be operated more dynamically according to the network load, e.g., some U-plane cells can be switched off if there is no user within their coverage area. Another option is to dedicate the macro-cell that works in the lower-frequency band to the control plane, while

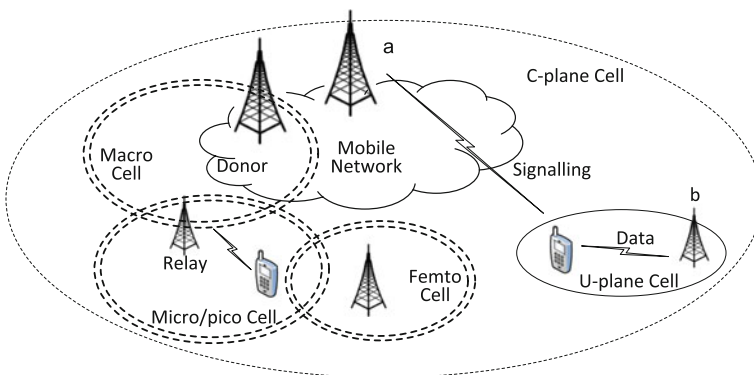


Fig. 9.2 HetNet and multilayered radio network

dedicating LPNs in higher-frequency bands to the user plane. From a spectrum band usage perspective, these architectures may be considered to be multilayered radio networks. This means that the control plane typically operates in one band, while the user planes can operate in different bands, and carrier aggregation (i.e., aggregation of carriers at different bands) may be used for pure data transmission purposes.

With regard to the HetNet and the novel multilayered radio network, the following requirements should be met from a mobility management viewpoint: low handover failure rates, avoidance of ping-pongs between different types of cells, and handover and location management enhancements for user movements among the novel radio network cells. When designing these mobility management enhancements for the deployment scenarios above, radio aspects such as interference, e.g., e-ICIC (enhanced intercell interference coordination), should also be taken into account.

(3) Machine-to-machine communication

Machine-to-machine communication (M2M, also known as machine-type communication, or MTC) is a new form of communication used in the following scenarios: surveillance systems for security purposes, metering systems for utility usage, tracking and tracing for animal control and logistics, remote maintenance and control of equipment such as pumps, lighting, and vending machines, health systems such as those used for monitoring of vital signs, and payment systems such as vending machines. Consumer devices such as tablets, smartphones, e-books, and digital photograph frames are also considered to be machines because they sometimes communicate like machine-type equipment, and their traffic patterns are similar to those of MTC communications.

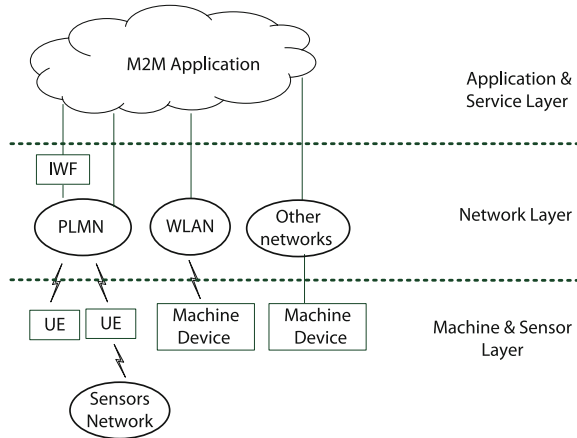
Compared with human-to-human communications, MTC has the following characteristics: No human is involved in the end-to-end communication, the number of devices involved may be huge, the quantity of data exchanged may be very small, the frequency of the communications may be frequent or infrequent, and the communication can be scheduled. Figure 9.3 shows a simplified MTC system. For PLMN, an interworking function has already been defined to transfer control plane-like messages from the M2M application to the user equipment (UE). User plane communication is also possible. The UE can also work as a gateway for other machine-type devices, such as sensors.

It should be noted that within the 3GPP specifications, UE is a general term used to refer to the UE used for MTC and/or the UE used by humans.

Each different Standards Development Organization (SDO) has a different focus, i.e., the European Telecommunications Standards Institute (ETSI) TC M2M and one M2M focused on the application and service layers, and 3GPP focused on the network layer and its UE, while other organizations such as IEEE 802.15.4 focused on sensor and machine devices.

Based on the above MTC scenarios, more specific requirements must be considered. From the application and service layer viewpoint, each M2M application

Fig. 9.3 Simplified MTC system



uses its own service enablement layer to communicate with the network. Whether a common layer or a unified service enablement layer is needed and how this layer will interwork with the network are hot issues that have been discussed recently. From the network-layer perspective, several issues have been identified: considering the huge numbers of MTC devices and the possible traffic that they could generate, and problems such as lack of identifiers and network congestion caused by massive network access should be considered. Other challenges stem from their different mobility patterns, e.g., some devices move quickly within a very large range, while some remain almost stationary; the traffic patterns are quite different from those of human communications, e.g., for metering, only very small amounts of data would be sent monthly, while the traffic for video surveillance would remain quite stable. From the terminal viewpoint, the cost of battery replacement is prohibitively high for some machine devices (e.g., for animal tracking); the related control plane and user plane overheads should be minimized as far as possible to prolong the life cycles of these devices.

9.2 Methodology-Related Challenges

9.2.1 Identifiers and Addressing in Future Internet

Naming and addressing are important issues in communication network architectures and include three concepts [9]: name (location independent), which indicates “what we seek”; the logical address (location information), which means “where it is”; and the route, which tells us “how to get there.” A mapping system exists between the logical address and the physical address. The naming and addressing rules reflect the mobility management capabilities of the network architecture. The identifier, which is unique and permanent, is a fundamental attribute for user access

to networks and services. The locator, which is both variable and temporary, is used to identify the host network access position. It is the foundation of mobility support and service continuity assurance.

In cellular mobile communication networks, each subscriber has a unique International Mobile Subscriber Identity (IMSI) that is valid in all network service domains. The IMSI is stored in the home location register (HLR) and the visitor location register (VLR) simultaneously, and it is used in call establishment and location update procedures. When the subscriber moves within the network coverage, the VLR available to the visiting users assigns a unique Temporary Mobile Subscriber Identity (TMSI) to each subscriber. The TMSI acts as the subscriber’s location identifier, which is valid in the VLR domain only. During call establishment and location updates, the TMSI is used instead of the IMSI when transmitted by the air interface for security reasons. Generally, cellular mobile communication networks achieve effective management of user mobility through the use of a pair of IMSI/TMSI identifiers, along with the HLR/VLR location mapping system.

In current Internet architectures, the IP address has dual semantics, which indicate both identity and location attributes. In the transport layer and the application layer, the IP address is used instead of an identifier to identify an end-to-end connection; in the network layer, the IP address is used instead of a locator for routing and addressing. Figure 9.4 shows the dual functionality of IP address.

Comparisons of IDs and locators in Internet Research Task Force Routing Research Group (IRTF-RRG) and International Telecommunication Union (ITU) standards are summarized in Table 9.4 [10, 11].

The semantic overload of the IP address means that TCP/IP-based networks are congenitally deficient in terms of mobility support capabilities. The Mobile IP (MIP) and its extensions use a pair of IP addresses, the home address (HoA) and the care-of address (CoA), and a mapping system have been used to solve the host mobility problem. The dynamic DNS (domain name system) uses dynamic updating methods to optimize the network support for mobility. However, in essence, these methods have failed to achieve any real breakthrough for IP mobility management, which has become a major bottleneck that is restricting future Internet development.

In recent years, the contradictions between the unique identifier and the dynamic locator have become the focus of the research on naming and addressing technology for the future Internet. The information-centric network (ICN), which represents a new network architecture, has attracted considerable attention. Designing models with separate identity attributes and location attributes has become an

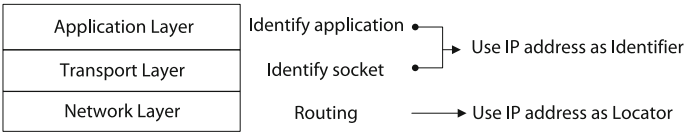


Fig. 9.4 Dual functionality of the IP address

Table 9.4 Comparisons of IDs and locators in IRTF-RRG and ITU standards

	ID	Locator
IRTF-RRG	Name of an object at a given layer	Name for a point of attachment within the topology at a given layer
ITU	Node ID used at the transport and higher layers to identify the node and the endpoint of a communication session	Topological name for an interface or a set of interfaces

inevitable trend. Basically, the ICN architecture supports consumer mobility and provider mobility effectively by designing an identifier that is independent of location. The consumer mobility is handled well because of the consumer-driven nature of most ICN architectures. In terms of provider mobility, the ICN must maintain routing consistency during provider movement. This is because it is clearly necessary to update the (global) locator information whenever a provider relocates.

In summary, we see that the future of the Internet naming and addressing technology still faces many problems and challenges, especially in the mobility support area with the definitions of identifiers and locators, the mapping system architectures, and the effects of new network architectures.

9.2.2 Emerging Radio Access Network Architectures

Mobility management in normal radio access networks is mainly focused on network selection for the idle mode and handover for connected modes. The same RAN nodes provide the related functionality for both the control plane and the user plane. Thus, the control plane and the user plane and their related procedures are integrated as a whole. This is a rather straightforward design. However, the decoupling of the C/U-plane poses great challenges for mobility management and other network areas. Figure 9.5 shows a new radio network architecture.

In the above architecture, a single centralized control plane node (which has control plane cells with large coverage) may have control links with multiple user plane nodes (X interfaces). The cells of the user plane nodes typically cover very small areas within the coverage of the control plane cell. The user plane-only cell has been called a virtual cell or phantom cell, or a small cell, where control-related functions such as cell-specific channels (e.g., cell-specific reference signal (CRS), master information block/system information block (MIB/SIB)) have been removed, but these functions are fulfilled by the control cell instead. The control plane node/cell is responsible for coordinating the radio resource management among the user plane nodes/cells that lie under its radio coverage, along with the RAN-level mobility procedure (e.g., idle model mobility, interuser plane node handover controls). To satisfy the high data rates required in some areas, virtual

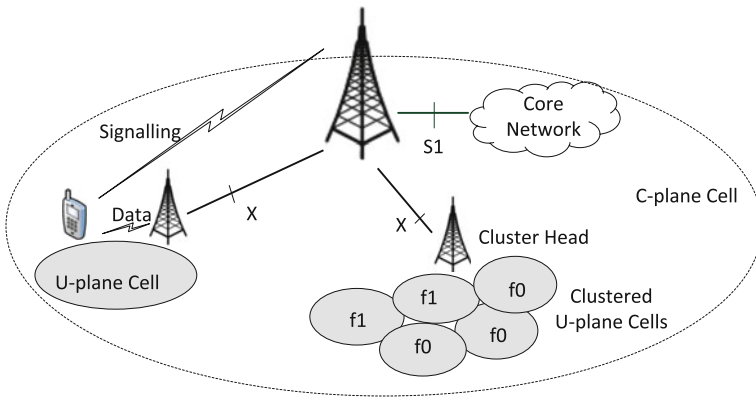


Fig. 9.5 Example of an emerging radio network architecture

cells in different frequency bands can be deployed as a single cluster, and one user plane node can serve as the cluster head, with a common user plane interface to the network for cells within the cluster.

The C/U-plane decoupling architecture is considered to be a major change from the network architecture perspective and will bring numerous challenges in mobility management. For idle mode mobility management, virtual cell discovery with low power consumption is required, and efficient mobility between virtual cells is another challenge. From the system viewpoint, and taking the LTE as an example, a new interface (and/or procedures) may possibly be needed between the C/U-plane nodes and the evolved packet system (EPS) core nodes, and this will result in additional functionality or modifications to the current mobility management schemes, assuming that there are no architectural changes in the LTE core network. Huge data volumes coming from massive U-plane cells will place a great burden on the current core network, and thus, a method to off-load the U-plane cell traffic presents another challenge. The most aggressive way for core networks to adapt to C/U-plane decoupling in RANs may lead to new core network architectures and to new schemes for the mobility management functions.

9.2.3 Mobility Management Function Structures: Centralized Versus Distributed

The mobility management architecture of centralized and distributed can be analyzed from two dimensions: control plane and data plane. They can be deployed in a centralized way, distributed way, or hybrid way. The function of control plane is to allocate IP address, manage, and keep track of location information of MNs, while the data plane is responsible for receiving and forwarding data packets

destined to a MN. Most mobility management technologies employ an anchor point to allow a MN to remain reachable after it has moved to a different network. The anchor ensures connectivity by forwarding packets destined to, or sent from, the MN. Generally, an anchor is defined as a combined control plane and data plane functions.

Centralized mobility management makes use of centrally deployed anchors, which refers to deploy very few anchors, and the traffic of millions of MNs is managed by the same anchor [12]. In centralized mobility management, the location information of a MN is kept at a single anchor, and packets destined to the MN are forwarded via this anchor. In other words, such mobility management is centralized in both control plane and data plane. However, centralized mobility management suffers from the lengthy path problem. For instance, centralized anchors manage all mobility contexts and routing for MNs, which can cause inefficient traffic routing because all traffic associated with the MNs is first routed to the centralized anchors and then forwarded to destinations. It thus affects network performance by slowing down the end-to-end transmission.

The explosion of mobile Internet traffic is accelerating the evolution of mobile network toward flat architectures with all-IP support. Distributed mobility management aims at distributing mobile Internet traffic in an optimal way while not relying on centrally deployed anchors. In distributed mobility management, anchors are distributed at the access network level so that a close by anchor is available to serve the MNs. Traffic associated with the MNs is naturally distributed and optimally delivered [13].

The centralized mobility management is applied in most cellular mobile networks, i.e., centralized signaling control and centrally deployed anchors. However, large amount of Internet traffic in the mobile networks causes heavy signaling cost and inefficient data forwarding. Therefore, in the future mobile network, decentralized network control is expected for signaling cost saving, and distributed gateway is desired for traffic off-loading. In a word, employing distributed mobility management becomes a trend.

Many existing IP mobility management protocols make use of centralized anchors that facilitate mobility support for all registered mobile nodes (MNs) in a hierarchical network. Examples are the home agent (HA) and local mobility agent (LMA) serving as the anchors for the MNs and Mobile Access Gateway (MAG) in MIPv6 and in PMIPv6, respectively. These protocols have been designed in a centralized architecture, where both the mobility context and the traffic encapsulation are maintained at a centralized anchor. No such a central anchor is deployed in distributed mobility management, but the anchors are distributed at multiple access routers (ARs). The distribution of anchors will help to avoid a centralized performance bottleneck and processes traffic associated with MNs in a distributed manner.

Some challenges for mobility management architecture include address and tunneling management, registration delay and signaling overhead, handover control and resource management, and security consideration.

9.2.4 Mobility Patterns

The mobility model represents the movements of mobile users and how their locations, velocities, and accelerations change over time [14]. It is an important basis for research in traditional mobility management technology and aims to accurately describe the mobility characteristics and models of mobile objects. It forms the premise for performance and overhead analysis of location registration, paging, and handover processes. It is an important factor that must be considered when designing mobility management mechanisms. Also, it provides a basis for proactive mobility management technology, which completes location management and handover functions based on efficient mobility prediction.

However, traditional mobility modeling uses the assumption of random movement and emphasizes the modeling of certain physical mobility parameters, such as location, velocity, and acceleration over time. It therefore cannot explore the

Table 9.5 Limitations of traditional mobility modeling and challenges in human mobility modeling

	Limitations of traditional mobility modeling	Challenges in human mobility modeling
Mobility metrics	Focuses on the changes in some basic physical mobility parameters, such as location, velocity, and acceleration over time	Requires multidimensional modeling in spatial, temporal, and social dimensions. Spatial metrics pertain to the behavior of the users in the physical space. Temporal metrics portray the time-varying features of human mobility. Social metrics describe the interactions between the users [15]
Assumptions	Often assumes user movement with random features	Recent research on human mobility indicates that human mobility does not follow random movement patterns in nature. It presents obvious regularity, periodicity, and predictability [16, 17]
Mobile object	Most mobility models (entity mobility models) depict the mobility of independent individual mobile nodes. Certain others (group mobility models) describe the mobility of a group of mobile nodes, but are limited to physically explicit groups	Also concerns the implicit groups formed according to social relationships
Scale	Small timescales, mobile nodes in specific scenarios	Large timescales, massive nodes
Modeling method	Stochastic process theory	Pattern mining, network science

mobility characteristics and laws comprehensively and thus cannot achieve the satisfactory optimization performance as expected.

Mobility patterns actually have wider meanings than traditional mobility models. These patterns concern the characteristics and regularities of the mobile objects in temporal, spatial, and social dimensions. Because mobile terminals are carried by people with social relationships, human mobility patterns have become an important basis for future research in mobility management technology. Exploring the inherent and essential characteristics and regularities of human mobility, using them for theoretical analysis of the network topology and network connectivity metrics and then applying them to protocol design and performance optimization, is the methodology for future research on mobility management.

Compared with the limitations of traditional mobility modeling, human mobility pattern modeling faces certain challenges, as shown in Table 9.5.

9.3 Solution Trends

9.3.1 *Design ID–Locator Architecture for Mobility*

The most important technology direction for the future Internet naming and addressing system is the combination of the uniform naming of the network resources with a rapid mapping mechanism. This aims to provide users with a location-independent identifier at the application layer, while the network-layer address serves as a locator.

The identifier/locator separation architecture can effectively solve many of the problems of current IP networks, including host mobility, multi-homing, routing scalability, and IPv4/IPv6 transitions. The mobile ICN is an ICN that supports, at some elements in the network path, consumer and/or provider mobility. Consumer mobility allows the consumers to reconfigure their network location without disrupting the connectivity, while provider mobility allows sources to relocate without disrupting the content availability.

The naming and addressing technology requirements for the future Internet include: (1) the definition of a unified naming method to support the distinctions between different types of moving targets, such as mobile terminals and mobile networks, (2) design of an efficient mapping system for user identifiers and locators to support fast information updates, thus ensuring communication continuity, and (3) more research into mobility management in ICNs, which is still at the initial stage. In addition to supporting host mobility, ICNs require the design of an effective mechanism to support server/provider mobility and thus ensure the network resource reachability during mobility.

9.3.2 Mobility Management Enhancements for Emerging Radio Access Network Architectures

Because the HetNet does not bring any changes to the core network, the current research focus involves enhancement of the existing radio network-level mobility management scheme while taking the HetNet into account. The overlapping coverage of the different cell types means that robust mobility management enhancements are required to achieve good handover success rates, avoid ping-pongs, and mitigate the interference caused by the multilayered cells.

For the emerging radio network architectures, both idle mode and connected mode management solutions or enhancements are needed. In the idle mode, the terminal usually moves across small cells frequently, and methods that enable the discovery of a suitable small cell when there is uplink data pending should be studied. Enhancement of the current location management functionality is also needed, e.g., for trade-offs between paging overheads and location accuracy. The possibility of massive paging messages being sent on a single control cell must also be studied in terms of aspects such as the capacity of the paging channel, the size and coding of the paging message, and ways to produce the paging response.

For connected mode mobility management, new procedures for handover control among small cells, virtual cells, and normal cells are required. Security-related mechanisms must also be enhanced for both idle mode mobility and connected mode mobility.

9.3.3 Distributed Dynamic Mobility Management

Distributed dynamic mobility management is a new approach for efficient mobility management within all-IP flat architectures to cope with the recent explosion of mobile Internet traffic. This solution will rely on mobility anchoring and on the dynamic update of the forwarding plane according to the location of the MN. The location information is maintained in a distributed manner.

The approaches of IETF DMM WG [18] may distribute anchors in flattening the mobility network such that the anchors are positioned closer to the user; ideally, anchors could be collocated with the first-hop router. Facilitated by the distribution of anchors, it may be possible to selectively use or not use mobility protocol support, depending on whether such support is needed or not [12].

By introducing distributed mobility management into cellular mobile network, the mobility management can be controlled by not only network function in core network but also localized network function in radio access network (RAN), and the anchor can be not only centralized packet data network (PDN) gateway but also

distributed local gateway. As a consequence, mobility can be efficiently handled by localized network function.

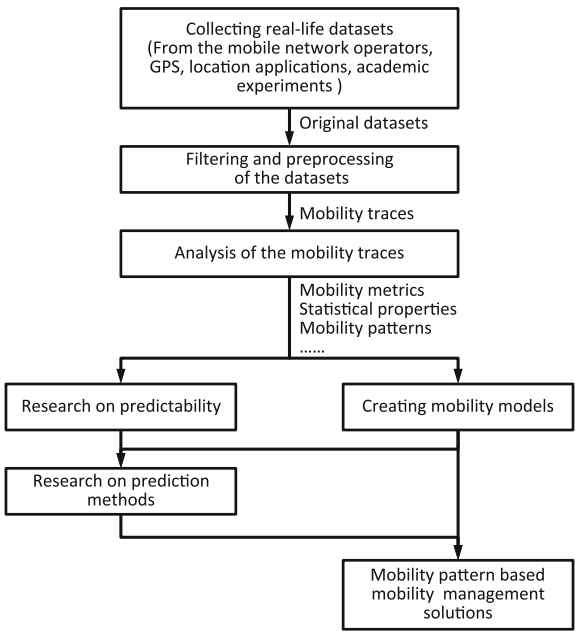
In the future mobile network supporting virtualization of network functions and separation of control plane and user plane functions, the functions for mobility management can be rapidly deployed on virtualized network infrastructure, and the functional entity can be placed flexibly, e.g., closer to RAN. By virtue of flexible deployment of mobility management functions, the distributed dynamic mobility management mechanism can be easily achieved.

9.3.4 Modeling of Mobility Patterns and Mobility Pattern-Based Mobility Management Solutions

As mentioned above, the mobility patterns define the mobility characteristics and laws in the temporal, spatial, and social dimensions. Accurate mobility pattern modeling provides an important basis for mobility management technology research. Figure 9.6 shows the general solution used in research for mobility pattern modeling and mobility pattern-based mobility management.

Firstly, modeling of mobility patterns that involve human behavior begins with the collection of real-life datasets, as shown in Fig. 9.6. These datasets can be

Fig. 9.6 General solution for mobility pattern modeling and mobility pattern-based mobility management



acquired from mobile network operators, GPS, certain location applications, or even the results of academic experiments. After the necessary filtering and pre-processing, some theoretical tools (e.g., physical statistical analysis, data mining, entropy, and complex networks) can be used for data analysis to acquire the mobility metrics, the statistical properties of these metrics, and the patterns that provide an important basis for the creation of mobility models.

Secondly, mobility prediction is based on the mobility patterns and laws. Mobility prediction in traditional mobility modeling is conducted based on pre-defined mobility models and physical kinematics and is limited to physical mobility metrics such as velocity, acceleration, and location. Although human mobility is indicated to have high predictability [17], efficient prediction mechanisms remain scarce in the existing research.

An efficient solution for mobility management optimization is to design schemes that conform to the mobility patterns and laws in nature. Possible solutions include the following:

- Design of location management areas based on mobility patterns. The location areas (LAs) and paging areas (PAs) should be designed based on the analysis of the mobility patterns to provide better trade-offs between efficiency, accuracy, and overheads in location management.
- Group mobility management based on group mobility behavior analysis. Exploration of the group mobility patterns, and identification of the explicit and implicit groups that are formed dynamically, will then allow the design of group mobility management solutions that conform to the mobility patterns.
- Radio resource management based on mobility patterns. For example, resource allocation may be conducted based on joint optimization of the physical topology in the actual network and the logical topology in the social network application. For another example, content-driven resource allocation can be designed based on the common interests and group mobility patterns of the users.

It is foreseen that mobility has become the intrinsic capability of the network and thus is envisioned to become the key driving force of the network evolution [19]. When prospecting the future networks, the emerging communication paradigms, such as mobile social network, mobile cloud computing, Internet-of-Things, and the expectation of ubiquitous and seamless connectivity, bring new requirements and new challenges for mobility management. It is necessary to clarify where mobility exists with a high-level abstraction, to abstract the key functions, entities, and principles from a systematic and unified perspective. Mobility-driven network (MDN) [20], as the vision of mobility management for the future network, follows the design philosophy of decoupling, including vertical decoupling of the mobility entities into service–entity and device–entity, and horizontal decoupling of the entity–identity and locator–identifier.

References

1. Kayastha N, Niyato D, Wang P et al (2011) Applications, architectures, and protocol design issues for mobile social networks: a survey. *Proc IEEE* 99(12):2130–2158. doi:[10.1109/JPROC.2011.2169033](https://doi.org/10.1109/JPROC.2011.2169033)
2. Zhong H, Bi L, Feng Z et al (2008) Research on the design method of mobile social network services. In: *IEEE international conference on information management, innovation management and industrial engineering, ICIII'08*, 2: 458–461. doi:[10.1109/ICIII.2008.206](https://doi.org/10.1109/ICIII.2008.206)
3. Zhang C, Yu C, Zhu X, Ya G (2012) *Technical foundations and development practices of SNS*. Posts & Telecom Press, Beijing
4. 3GPP TR23.703 (2013) Study on architecture enhancements to support proximity-based services (ProSe), Ver 1.0.0
5. GPP TR36.839 (2012) Evolved universal terrestrial radio access (E-UTRA); mobility enhancements in heterogeneous networks
6. Niu Z, Zhou S, Zhou S, Zhong X, Wang J (2012) HCN: energy efficiency and resource optimized hyper-cellular mobile communication system architecture and its technical challenges. *Sci China Inf Sci* 42(10):1191–1203
7. Ishii H, Kishiyama Y, Takahashi H (2012) A novel architecture for LTE-B: C-plane/U-plane split and Phantom cell concept. *IEEE Globecom Workshops (GC Wkshps)* 2012:624–630. doi:[10.1109/GLOCOMW.2012.6477646](https://doi.org/10.1109/GLOCOMW.2012.6477646)
8. GPP TR36.932 (2013) Scenarios and requirements for small cell enhancements for E-UTRA and E-UTRAN
9. Shoch J (1978) Inter-network naming, addressing, and routing. *IEEE Computer Society*, Fall, COMPCON
10. IETF RFC6306 (2011) Hierarchical IPv4 framework
11. ITU-T Y.2015 (2009) General requirements for ID/locator separation in NGN
12. IETF RFC7333 (2014) Requirements for distributed mobility management
13. Lee J-H, Bonnin J-M, Seite P et al (2013) Distributed IP mobility management from the perspective of the IETF. *IEEE Wirel Commun* 20(5):159–168. doi:[10.1109/MWC.2013.6664487](https://doi.org/10.1109/MWC.2013.6664487)
14. http://en.wikipedia.org/wiki/Mobility_model. Accessed 25 Mar 2014
15. Karamshuk D, Boldrini C, Conti M et al (2011) Human mobility models for opportunistic networks. *IEEE Commun Mag* 49(12):157–165. doi:[10.1109/MCOM.2011.6094021](https://doi.org/10.1109/MCOM.2011.6094021)
16. Gonzalez MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. *Nature* 453(7196):779–782
17. Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. *Science* 327(5968):1018–1021. doi:[10.1126/science.1177170](https://doi.org/10.1126/science.1177170)
18. <http://datatracker.ietf.org/wg/dmm/>. Accessed 29 Dec 2015
19. Raychaudhuri D (2013) Cellular-internet convergence: evolving the internet architecture to support mobility services as the norm. http://mobilityfirst.winlab.rutgers.edu/documents/Johannesburg_Summit_talk_513R.pdf. Accessed 15 Apr 2014
20. Chen SZ, Shi Y, Hu B, Ai M (2014) Mobility-driven networks (MDN): from evolution to visions of mobility management. *IEEE Netw* 28(4):66–73

Index

A

Access control, 8, 108, 126, 139, 190, 235
Access link, 10, 64, 71, 182, 212, 213
Access point (AP), 16, 37, 66, 126, 141, 146, 160, 228, 230, 269, 271
Addressing, 148, 209, 223, 246, 295, 296, 301
All-IP network, 135, 136
Attachment point, 18, 202, 215, 217, 241, 280, 288, 289
Authentication, 4, 5, 10, 24, 34, 36–39, 59, 60, 70, 76, 78, 127, 129, 130, 135, 139, 146, 158, 161, 179–181, 183, 185, 186, 188, 191, 197, 198, 232, 234–237, 261, 262, 266, 268, 272, 275–277, 280, 292
Authentication, authorization and accounting (AAA), 34, 39, 53, 57, 78, 135, 189–191, 195, 197, 198, 232, 234, 235, 237, 272, 273

B

Base station (BS), 4, 48, 66, 95, 107, 112, 122, 124, 125, 129, 131, 146, 147, 180, 182, 183, 189, 223, 271, 274, 287, 293

C

Cell, 4, 18, 37, 38, 44, 46–48, 50, 51, 56, 57, 92, 94–96, 100–102, 106, 107, 111, 112, 121, 124, 125, 130, 137, 139, 143–147, 149–155, 161, 164, 166, 167, 169, 170, 173–176, 271, 292–294, 297, 302
Circuit switch/packet switch, 122, 123, 130, 132, 135, 166, 175
Confidentiality, 232, 262, 268, 275, 276
Context, 78, 79, 82, 106, 141, 146, 148, 153, 155, 157, 158, 160–162, 167, 168, 174, 175, 178, 179, 186, 230, 233, 251, 258, 272, 286, 293, 299
Control plane, 29, 33, 57–60, 132, 139, 140, 142, 157, 168, 220, 292, 294, 295, 297, 298

D

DAR extension, 59, 246, 249, 251, 252, 255, 261
Data origin authentication, 232
Distributed mobility management, 20, 292, 299, 302
Downlink, 47, 84, 149, 156, 157, 160, 162, 168, 170, 172, 173, 227, 230

F

Fast handover, 15, 47, 58, 228–230, 272
Fault, configuration, accounting, performance, and security (FCAPS), 34
Foreign agent (FA), 8, 32, 37, 38, 41, 54, 57, 71, 195, 197, 203, 204, 206, 209, 223–226, 228, 233, 235
Foreign link, 204, 207
Foreign network, 42, 203, 204, 206, 207, 209, 211, 216, 217, 222, 227
Fuzzy logic theory, 92, 110

G

Group mobility, 19, 97, 99, 100, 288, 289, 300, 304
Global system for mobile communications (GSM), 5, 16, 18, 24, 36–38, 46, 50, 57, 66, 68, 84, 121, 123, 129–132, 135, 137, 139, 140, 147, 149, 150, 152, 154, 155, 179–185, 191

H

H.323, 41, 265
Handover, 4, 5, 10–12, 15–18, 20–22, 29, 31–33, 37, 38, 46–48, 50, 52–55, 57, 64–66, 68, 70, 71, 81, 93, 95, 96, 100, 102, 103, 105, 109, 111, 112, 124, 125, 128–130, 133, 137, 138, 146, 147, 149–151, 155, 160, 164–168, 170, 172, 173, 175, 176, 178, 185, 186, 188, 190,

192, 194, 195, 197, 211, 214, 221, 223, 227, 228, 242–246, 251, 252, 254–258, 260, 272, 273, 279–281, 292, 294, 297, 302

Handover control, 5, 10, 17, 22, 33, 37, 38, 45, 48, 49, 51, 52, 54, 57–60, 67, 91, 93, 105, 107, 111, 138, 139, 141, 143, 164, 165, 175, 177, 178, 189, 213, 221, 227, 255–258, 265, 266, 268, 271, 297, 299, 302

Handover decision, 6, 48–50, 67–69, 71, 72, 92, 95, 103, 104, 108, 109, 111, 137, 146, 258, 260

Handover delay, 47, 51, 70, 92, 107, 112, 146, 186, 227, 229, 272

Home agent (HA), 8, 32, 35, 37–39, 41, 54, 57, 58, 64, 66, 71, 80, 82, 135, 192, 193, 196, 202–205, 207–211, 214, 216–223, 225–227, 232–236, 253, 254, 280, 299

Home link, 193, 198, 202–204, 236

Home location register (HLR), 8, 36–38, 41, 56, 57, 76, 130, 141, 153, 175, 184, 189, 296

Home network, 12, 32, 35, 37, 38, 56, 75, 130, 133, 192, 197, 202–204, 207, 208, 211, 216, 235, 238, 280

Home network prefix, 64, 193, 212, 214, 216

Horizontal handover, 6, 15, 46, 66, 108, 112

Horizontal mobility, 11, 14

Host identity protocol (HIP), 6, 20, 31, 65, 73

Host mobility, 5, 12, 201, 202, 211, 214, 272, 296, 301

I

Identifier, 13, 20, 31, 56, 73, 148, 158, 160, 161, 167, 168, 170, 180, 189, 218, 220, 230, 231, 249, 268, 271, 286, 295, 296, 301, 304

Inter-AN mobility, 11, 13, 14

Inter-CN mobility, 13, 14

Interface, 5, 23, 38, 52, 70–72, 76, 77, 107, 122, 124, 129–133, 135, 139, 141, 147, 151, 154, 156, 166, 168, 174, 175, 180, 185, 186, 189, 193, 197, 201, 202, 213, 216, 217, 219, 242, 243, 245, 250, 258, 286, 293, 296, 297

Interoperability control, 10, 29, 33, 52, 91

Intra-AN mobility, 13, 14

IP flow, 19, 194, 198, 201, 202, 219, 220, 292, 293

IP flow mobility, 194

IP paging, 221, 222, 224, 227

IP routing, 31, 204, 224

IP session continuity, 195–197

L

Latency, 4, 19, 45, 166, 227, 228, 244, 247, 260, 261

Location area, 5, 35, 40, 41, 93, 128, 138, 141, 151–155, 161, 175, 221, 222, 304

Location database, 5, 40, 42, 43, 45, 92, 94, 107, 224, 253, 254

Location management, 4, 5, 10, 15–18, 22, 29, 33, 37, 38, 40, 44, 54, 57–60, 91, 93–95, 100, 102, 105, 107, 113, 130, 137, 138, 141, 143, 151–153, 158, 189, 212, 221–223, 242–248, 252, 254, 265, 266, 268, 269, 280, 292, 294, 300, 302, 304

Location update, 4, 5, 16, 37, 40, 41, 43, 44, 54, 57, 58, 65, 80, 92–94, 96, 100, 102, 107, 154, 160, 162, 176, 178, 188, 221, 224, 232, 272, 296

Locator, 6, 20, 31, 73, 296, 301, 304

Locator/identifier separation protocol (LISP), 20, 74

Long term evolution (LTE), 17, 18, 24, 122, 127, 128, 132–134, 137, 139–142, 144, 146, 147, 149–151, 153–155, 157, 158, 160, 161, 163, 166, 170, 175, 177, 179, 180, 182–186, 198, 290, 292, 293, 298

M

Macro-mobility, 5, 11, 13, 31, 64, 99, 222, 224

Management information base (MIB), 35, 297

Management plane, 29, 34, 139, 220

Markov process, 92, 95, 101, 103

Media independent handover (MIH), 23, 71, 272

Micro-mobility, 5, 11, 13, 31, 222, 224

MIPv4, 20, 32, 37–39, 41, 54, 57, 58, 65, 71, 195, 197, 203–205, 208, 221, 222, 224, 228, 233, 235

MIPv6, 20, 35, 39, 57, 58, 65, 71, 80, 81, 84, 85, 197, 205–208, 214, 216, 217, 219, 221, 222, 224, 228, 230, 233, 234, 236, 280, 299

Mobile assisted handover (MAHO), 49, 50, 52, 70, 147, 258

Mobile controlled handover (MCHO), 49, 50, 52, 70, 147, 258

Mobile host (MH), 16, 57, 107, 113, 202, 216, 224, 245, 246, 252, 253, 268, 278

Mobile internet protocol (MIP), 3, 5, 32, 38, 41, 48, 50, 52, 54, 55, 57, 58, 63, 70, 202, 203, 211, 214, 219, 221, 224, 233, 235, 245, 246, 252, 254, 296

Mobile node (MN), 13, 16, 20, 31, 32, 54, 63–65, 79, 80, 92, 97, 202, 203, 209, 214, 219, 221, 241, 269, 288, 299, 300

Mobile objective, 3, 34, 35, 39–42, 45, 300
 Mobile service, 10, 39, 285, 286
 Mobile social network (MSN), 286–289
 Mobile station (MS), 4, 16, 37, 38, 47–50, 102, 112, 122
 Mobile terminal (MT), 5, 16, 44, 48, 52, 54, 67, 69, 92, 96, 100, 102–104, 146, 151, 153–156, 164–166, 168, 175, 178, 266, 286
 Mobile user, 4, 17, 39, 41, 43–45, 58, 75, 79, 94, 96, 99, 128, 141, 243, 286, 287, 300
 Mobility, 3, 5, 7, 10–12, 14, 16–18, 20, 21, 23, 32, 33, 38, 39, 41, 45, 49, 52–54, 64, 72, 75, 77, 80, 81, 92, 93, 95, 96, 99, 102, 107, 113, 126, 127, 129–132, 135, 137, 139, 143, 145, 157, 164, 174, 185, 186, 202, 219, 233, 241
 Mobility anchor, 20, 302
 Mobility management, 3, 5, 7, 8, 10, 14, 16, 18, 19, 21, 24, 25, 29, 31, 33, 35, 38, 39, 45, 48, 52–55, 57–59, 60, 63, 65, 66, 73–75, 78, 80, 81, 91, 93, 95, 96, 101, 105–108, 111, 121, 122, 126, 128, 130–135, 137, 139, 141, 145, 152, 158, 177, 179, 181, 185, 188, 189, 195, 196, 198, 211, 212, 214, 216, 220, 222–224, 230–232, 241, 244–246, 249, 252, 254, 265, 267, 277, 285, 286, 289, 291, 292, 294–299, 301–304
 Mobility models, 6, 96, 99, 301, 304
 Mobility pattern, 45, 98, 141, 153, 154, 231, 287, 288, 295, 300, 301, 303, 304
 Mobile stream control transmission protocol (MSCTP), 6, 20, 33, 54, 59, 70, 241, 245–247, 249, 251–255, 257, 258, 260, 261
 Multi-homing, 31, 73, 245, 246, 249, 250, 252, 255, 260
 Multiple IP mobility, 197
 Multiplexing, 122, 123, 140

N

Network controlled handover (NCHO), 49, 52, 70, 147, 258
 Network coverage, 151, 290, 296
 Network interface, 13, 48, 53, 67, 72, 136, 203, 249, 255, 273
 Network mobility (NEMO), 5, 9, 11, 12, 20, 25, 31, 32, 63, 79–85, 201, 202, 214, 215, 217, 288, 289
 Network reference model, 8, 25, 35, 56, 57, 136, 137

P

Packet domain, 132, 133
 Packet switched domain, 122
 Paging, 4, 5, 16, 35, 37, 38, 40, 92–94, 96, 101, 102, 107, 113, 130, 134, 137, 139, 144, 145, 151, 155–157, 175, 221, 225, 227, 271, 300, 302
 Personal mobility, 9, 11, 12, 25, 31–33, 54, 59, 63, 74, 75, 277, 278
 PMIPv6, 20, 64, 65, 81, 82, 85, 195, 197, 202, 211–214, 226, 227, 229, 231–233, 237, 299
 Point-to-point link, 198
 Prefixes match, 193
 Probability theory, 25, 91–93
 Protocol reference model, 8, 25, 29, 33, 35, 57, 59, 60, 136, 137, 139, 292
 Proxy server, 54, 60, 71, 254, 266–269, 275, 280

Q

Queuing theory, 25, 91, 92, 95, 105–108

R

Redirect server, 54, 60, 71, 266–269, 276, 280
 Reference point, 100, 195–197, 292
 Registrar, 54, 60, 71, 254, 266–268, 270, 276, 277, 280
 Registration, 4, 5, 32, 36–39, 54, 57–60, 70, 76, 102, 113, 135, 137, 141, 143–145, 153–155, 157, 158, 160, 161, 178, 186, 193, 202, 205, 206, 208, 216, 219, 221, 223, 225–228, 231, 233, 235, 236, 261, 268–270, 272, 275, 280, 281, 299
 Relocation, 147, 163, 167–169, 173
 Roaming, 4, 5, 12, 21, 23, 24, 36, 53, 78, 122, 126, 128–130, 133, 176, 189, 192, 195, 198, 234
 Routing area, 138, 141, 145, 152, 153, 155, 156, 168, 169, 176

S

SCTP association, 54, 250, 251, 253–255, 258, 262
 SCTP endpoint, 246, 249, 250, 258
 Seamless handover, 15, 19, 23, 47, 139, 228, 245, 246
 Seamless mobility, 11, 15, 17, 21, 53, 122, 197
 Security association, 58, 192, 193, 232, 262
 Security management, 29, 34, 58, 60, 76, 141, 182
 Security tunnel, 192, 195, 196

Service continuity, 11, 18, 22, 25, 132, 134, 135, 165, 185, 188–190, 192, 202, 208, 291, 292, 296

Service mobility, 9, 11, 12, 25, 31–33, 54, 59, 63, 77, 78, 279

Session continuity, 10, 12, 15, 23, 33, 34, 45, 66, 71, 72, 74, 84, 138, 219

Session initiation protocol (SIP), 6, 20, 31, 33, 37, 41, 48, 54, 55, 59, 70, 76, 78, 107, 246, 252, 254, 265–277, 279–281

SHIM6, 31

Smooth handover, 15, 46, 58, 228

Social, local, and mobile (SoLoMo), 286, 287, 290

Soft handover, 15, 46–48, 51, 57, 71, 106, 112, 131, 149, 243, 245–247, 253, 268, 272–275, 281

Stochastic process, 25, 91, 92, 95, 96, 101, 300

T

Terminal mobility, 9, 11, 12, 20, 25, 31, 32, 54, 59, 63, 74, 80, 194, 195, 277

Transport plane, 220

Triangle routing, 204, 205, 208

Tunnel, 38, 64, 66, 80, 82, 132, 141, 146, 149, 156, 160, 161, 166, 167, 170, 172, 173,

186, 192, 193, 198, 202, 210, 212–214, 216–218, 228–232, 234

U

Uplink, 47, 149, 160, 161, 167, 168, 170, 172, 173, 231, 302

User agent (UA), 60, 267, 273, 275, 276, 280

User plane, 124, 132, 138, 139, 141, 142, 146, 150, 156, 162, 166, 168, 173, 182, 183, 193, 195, 223, 292–295, 297, 298, 303

V

Vertical handover, 6, 15, 23, 46, 48, 49, 66–72, 92, 101, 103, 104, 108–111, 227, 258, 260

Vertical mobility, 11, 14

Visitor location register (VLR), 8, 37, 41, 56, 57, 130, 138, 141, 152, 153, 155, 156, 160, 162, 164, 180, 182, 184–186, 296

W

Wireless local area network (WLAN), 13, 16, 31, 37, 46, 50, 53, 66, 68, 79, 103, 109, 111, 121, 122, 126, 127, 132, 135, 188–193, 195, 198, 258, 291, 292