Abd-Elhamid M. Taha | Najah Abu Ali | Hossam S. Hassanein

LTE, LTE-Advanced and WiMAX

Towards IMT-Advanced Networks

WILEY

LTE, LTE-ADVANCED AND WiMAX

LTE, LTE-ADVANCED AND WiMAX

TOWARDS IMT-ADVANCED NETWORKS

Abd-Elhamid M. Taha and Hossam S. Hassanein
Both of School of Computing, Queen's University, Canada

Najah Abu Ali
College of Information Technology, UAE University, United Arab Emirates

WILEY
A John Wiley & Sons, Ltd., Publication

This edition first published 2012

Registered office
John Wiley & Sons Ltd., The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data
Hassanein, H. (Hossam)
LTE, LTE-advanced, and WiMAX : towards IMT-advanced networks / Hossam S. Hassanein, Abd-Elhamid M. Taha, Najah Abu Ali. – 1st ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-74568-7 (hardback)
1. Long-Term Evolution (Telecommunications) 2. IEEE 802.16 (Standard) I. Taha, Abd-Elhamid M. II. Ali, Najah Abu. III. Title.
TK5103.48325.H37 2012
621.3845′6 – dc23

2011025964

A catalogue record for this book is available from the British Library.

Print ISBN: 9780470745687
ePDF ISBN: 9781119970453
oBook ISBN: 9781119970446
ePub ISBN: 9781119971467
mobi ISBN: 9781119971474

Set in 10/12pt Times by Laserwords Private Limited, Chennai, India.

To the memory of Mohamed Taha,
and the great father he was.
Abd-Elhamid

To my family, with a gratitude deep
beyond what words can express.
Najah

To my loving family.
Hossam

Contents

About the Authors

Abd-Elhamid M. Taha holds a strong expertise in wireless access technologies and networks. He has written and lectured on the subject of broadband wireless networks, with special emphasis on the design and deployment of radio resource management frameworks. He is currently a researcher and an adjunct assistant professor at Queen's University, Kingston, Ontario.

Najah Abu Ali is an expert on Access Wireless Networks architecture, design, QoS provisioning, implementation and performance. Her research interests comprise wired and wireless communication networks. Dr. Abu Ali has published and lectured widely on the subject of broadband wireless networks and their enabling technologies.

Hossam Hassanein is a leading authority in the areas of broadband, wireless and mobile networks architecture, protocols, control and performance evaluation. His record spans more than 300 publications in journals, conferences and book chapters, in addition to numerous keynotes and plenary talks in flagship venues. He is also the founder and director of the Telecommunications Research (TR) Lab at Queen's University School of Computing, with extensive international academic and industrial collaborations. Dr Hassanein is an IEEE Communications Society Distinguished Lecturer.

Preface

This is a book about IMT-Advanced access networks.

It is also a book that describes how these networks will be able to satisfy the ever increasing demand for mobile data. By some estimates, mobile traffic will take up to 6.3 exabytes (that is, 6.3 mega terabytes) per month in 2015. In 2015, there will also be one mobile device per capita – something in the range of 7.2 to 7.5 billion devices connected to the a wireless network. In 2020, the number of connected wireless devices will be more than 50 billion.

In 2008, the International Telecommunications Union – Radio Communications Sector (ITU-R) issued the requirements for the next generation cellular networks. In the requirements, the ITU-R the goals for the performance requirements of IMT-Advanced networks. The goals were ambitious relative to their predecessors, IMT-2000 or 3G networks, but not in terms of technologies. Simply put, the requirements had to do with accommodating the above noted increasing demand. They also had to do with enhancing the user overall wireless experience, starting from reducing the cost of the mobile handset the wireless device; reducing the cost and enhancing the quality mobile access; providing better support for both indoors and outdoors, in addition to higher quality connections at different mobility speeds. The requirements also made better international roaming a mandate. For operators, the requirements facilitated economic deployment, expansion and operation of wireless networks – a highly sought objective, especially after the great investments that were made in 3G networks.

In October 2010, the ITU-R recognized 3GPP's LTE-Advanced and IEEE's 802.16m (WiMAX 2.0) as two technologies satisfying the requirements for next generation wireless.

This book describes the technologies and functionalities that are enabling the two standards to realize these requirements. The exposition adopted parts from the traditional ways in which the two standards are introduced, which have generally been to follow the outlines of their respective recommendations. Instead, this book takes a "functionality-based" view, discerning information that answer questions like "what's IEEE 802.16m relay frame structure like?", "how does a UE camp on an LTE-Advanced cell?" or "how is security different in WiMAX

from LTE?" This view, while more tiresome to develop, makes it easier for the practitioner and the researchers to get to the heart of things quickly and with ease.

Our hope is that you will find our efforts useful.

Abd-Elhamid M. Taha
Najah Abu Ali
Hossam S. Hassanein

Acknowledgements

This book would not have been possible if it wasn't for the support of many.

The great (and very patient) editorial staff of Wiley & Sons, including Mark Hammond, Sarah Tilley, Sophia Travis, Susan Barclay, Mariam Cheok, and Keerthana Panneer of Laserwords Private Limited. Thank you for facilitating this book and making it possible.

The Broadly Project students at the Telecommunications Research Lab at the School of Computing, Queen's University, including (by alphabetical last name) Hatem Abou-Zeid, Hassan Ahmed, Abdallah Almaaitah, Mervat Fahmy, Pandeli Kolomitro, Mahmoud Ouda, Samad Razaghzadah, Mohamed Salah, and Nassif Shafi. Thank you for helping out at various parts of this book's development.

Ala Abu Alkheir did an excellent job in providing a much valued review of several chapters towards final stages of writing this book.

Sam Aleyadeh put in a lot of effort throughout into preparing the book's artwork, in addition to overseeing the required permissions from both IEEE and 3GPP.

Finally, we acknowledge the constant support of our families – one that was provided in many uncountable ways. We can never thank you enough.

List of Abbreviations

1G	First Generation Wireless Networks
2G	Second Generation Wireless Networks
3G	Third Generation Wireless Networks
3GGP2	Third Generation Partnership Project 2
3GPP	Third Generation Partnership Project
4G	Fourth Generation Wireless Networks
AAA	Authentication, Authorization and Accountability
ABS	Advanced Base Station
ACK	Acknowledgement message
ACM	Adaptive Coding and Modulation
ADC	Analog to Digital Conversion
AF	Amplify and Forward
AKA	Authentication and Key Agreement
A-MAP	Advanced allocation map
AMBR	Aggregate Maximum Bit Rate
AMS	Advanced Mobile Subscriber/Station
AN	Access Network
ARQ	Automatic Repeat Request
ARS	Advanced Relay Station
AS	Access Stratum
ASN	Access Service Network
ATM	Asynchronous Transfer Mode
BCCH	Broadcast Control Channel
BCH	Broadcast Channel
BE	Best Effort
BER	Bit Error Rate
BR	Bandwidth Request
BS	Base Station
BSID	Base Station ID
CAC	Call Admission Control
CBR	Constant Bit Rate
CCCH	Common Control Channel

CDMA	Code Division Multiple Access
CDS	Channel Dependent ascheduling
CGI	Cell Global Identification
CI	Cell Identifier
CID	Connection ID
CINR	Carrier-to-Interference-and-Noise-Ratio
CMAS	Commercial Mobile Alert System
CoMP	Coordinated Multipoint Transmission
CP	Cyclic Prefix
CPS	Common Part Sublayer
CQI	Channel Quality Indicator
CQICH	Channel Quality Indicator Channel
CRC	Cyclic Redundancy Check
C-RNTI	Cell Radio Network Temporary Identifier
CS	Service Convergence Sublayer
CSG	Closed Subscriber Group
CSI	Channel State Information
CSN	Connectivity Service Network
DAC	Digital to Analog Conversion
DBPC-REQ	Downlink Burst Profile Change Request
DBPC-RSP	Downlink Burst Profile Change Response
DCCH	Dedicated Control Channel
DCD	Downlink Channel Descriptor
DeNB	Donor eNB
DFT	Discrete Fourier Transformation
DHCP	Dynamic Host Configuration Protocol
DL	Downlink
DL-MAP	Downlink allocation map
DL-SCH	Downlink Shared Channel
DOCSIS	Data Over Cable Service Interface Specification
DRR	Defict Round Robin
DRX	Discontinuous Reception
DSA	Dynamic Service Addition
DSA-REQ	Dynamic Service Addition Request
DSA-RSP	Dynamnic Service Addition Response
DSC	Dynamic Service Change
DSC-REQ	Dynamic Service Change Request
DSC-RSP	Dynamic Service Change Response
DSD	Dynamic Service flow Deletion
DwPTS	downlink part
EAP	Extensible Authentication Protocol
EDF	Earliest Deadline First
EDGE	Enhanced Data Rates for GSM Evolution
eNB	enhanced Node B

EPC	Evolved Packet Core
EPS	Evolved Packet System
ertPS	extended real time Polling Service
ETWS	Earthquake and Tsunami Warning System
EUTRAN	Evolved Universal Mobile Telecommunications System
EVDO	Evolution-Data Optimized
EXP/PF	exponential/proportional fair
FBSS	Fast Base Station Switching
FCH	Frame Control Header
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
Femto ABS	Femtocells Advanced Base Station
FID	Flow ID
FIFO	First Input First Output
FR	Frequency Reuse
FTP	File Transfer Protocol
FUSC	Full Usage Subcarrier
GBR	Guaranteed Bit Rate
GERAN	GSM EDGE Radio Access Network
GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio Service Network
GSA	Global mobile Suppliers Association (www.gsacom.com)
GSM	Global System for Mobile Communications
GSMH	Grant Management Subheader
GTP	GPRS Tunneling Protocol
GUTI	Globally Unique Temporary Identity
H(e)MS	HeNB Management System
HARQ	Hybrid/Automatic Repeat Request
HeMS	HeNB Management System
HeNB	Home eNB
HeNB-GW	Gateway HeNB
H-FDD	Half-Frequency Division Duplex
ICI	Inter-Carrier Interference
IDFT	Inverse Discrete Fourier Transformation
IE	Informationa Element
IEEE	Institute of Electric and Electriconic Engineers
IETF	Internet Engineering Task Force
IFFT	Inverse Fast Fourier Transformation
IMEI	International Mobile Station Equipment Identifier
IMSI	International Mobile Subscriber Identity
IMT	International Mobile Telecommunications
IMT-2000	International Mobile Telecommunications – 2000 (or 3G)

IMT-Advanced	International Mobile Telecommunication – Advanced
IP	Internet Protocol
IPSec	Internet Protocol Security
ISI	Inter-Symbol Interference
ITU	International Telecommunications Union
ITU-D	International Telecommunications Union – Development Section
ITU-R	ITU Radiocommunications Sector
LOS	Line of Sight
LTE	Long Term Evolution
LTE-Advanced	Long Term Evolution Advanced
MAC	Medium Access Control
MAD	Minimum Area Difference
MBMS	Multimedia Broadcast and Multicast Services
MBR	Maximum Bit Rate
MCS	Modulation and Coding Schemes
MDHO	Macro-Diversity Handover
MIB	Management Information Base or Master Information Block
MIH	Media Independent Handover
M-LWDF	Maximum-Largest Weighted Delay First
MME	Mobile Management Entity
MOB_ASC_REPORT	Association Report message
MOB_BSHO-REQ	Base Station Handover Request message
MOB_BSHO-RSP	Base Station Handover Response message
MOB_HO-IND	Handover Indication message
MOB_MSHO-REQ	Mobile Station Handover Request message
MOB_MSHO-RSP	Mobile Station Handover Response message
MOB_NBR-ADV	Neighbor Advertisement message
MOB_SCN-REQ	Scanning Interval Allocation Request message
MOB_SCN-RSP	Scanning Interval Allocation Response message
MR	Multihop Relay
MR-BS	Multihop Relay Base Station
MRS	Mobile Relay Station
MS	Mobile Subscriber/Station
NACK	Negative Acknowledgement message
NAS	Non Access Stratum
NCMS	Network Control and Management Systems
NDI	New Data Indicator
NLOS	Non-Line of Sight
nrtPS	non real time Polling Service
ntRS	non-transparent Relay Station
OECD	Organization for Economic Cooperation and Development

OFDMA	Orthogonal Frequency Division Multiple Access
OSG	Open Subscriber Group
PAPR	peak to average power ratio
PBCH	Physical Broadcast Channel
PCCH	Paging Control Channel
PDCCH	Packet Data Control Channel
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
PDU	Protocol Data Units
PF	Proportional Fair
PF-BST	Proportional Fair Binary Search Tree
P-GW	PDN Gateway
PHICH	Physical HARQ Indicator Channel
PHY	Physical Layer
PLMN	Public Land Mobile Network
PM bit	Poll Me bit
PMP	Point to Multi-Point
PRB	Physical Resource Block
PSS	primary synchronization signal
PSTN	Public Switched Telephone Network
PTI	Procedure Transaction ID
PUSC	Partial Usage Subcarrier
PUSCH	Physical uplink Shared Channel
QCI	QoS class Identifier
QoS	Quality of Service
R1 BS	Legacy (IEEE 802.16-2009) Base Station
R1 MS	Legacy (IEEE 802.16-2009) Mobile Station
R1 RS	Legacy (IEEE 802.16-2009) Relay Station
RACH	Random Access Channel
R-ACK	Relay Acknowledgement
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Resource Block
RC	Resource Chunck
RCID	Reduced CID
RIT	Radio Interface Technology
RLC	Radio Link Control
R-MAP	Relay allocation map
R-NAK	Relay Negative Acknowledgement
RNG	Ranging
RNG-REQ	Ranging Request
RNG-RSP	Ranging Response
RNTI	Radio Network Temporary Identifier
R-PDCCH	Relay-Physical Downlink Control Channel

R-PDSCH	Relay-Physical Downlink Shared Channel
R-PUSCH	Relay-Physical Uplink Shared Channel
RRC	Radio Resource Control
R-RTG	Relay RTG
RS	Relay Station
RS-SCH	RS Scheduling information
RTG	Receiver-Transmitter Transition Gap
rtPS	real time Polling Service
R-TTG	Relay STG
RV	Redundancy Version
Rx	Receiver
SAE	System Architecture Evolution
SAP	Service Access Point
SBC	SS Basic Capability
SBC-REQ	SS Basic Capability Request
SBC-RSP	SS Basic Capability Response
SC-FDMA	Single Carrier Frequency Division Multiple Access
SCTP	Stream Control Transmission Protocol
SDF	Service Data Flow
SDU	Service Data Unit
SeGW	Security Gateway
SFH	Superframe Header
SFID	Service Flow ID
SFN	System Frame Number
SGSN	Serving GPRS Support Node
S-GW	Serving Gateways
SIB	System Information Block
SIB1	System Information Block Type1
SIB2	System Information Block Type2
SIB3	System Information Block Type3
SLA	service level agreement
SMS	Short Messaging Service
SN	Serving Network
SNMP	Simple Network Management Protocol
SNR	Signal to Noise Ratio
SRB	Signal Radio Bearer
SSS	Secondary Synchronization Signal
SSTTG	SS Transmission Receive Roundtrip Gap
STID	Station ID
STR	Simultaneous Transmit and Receive
TAC	Tracking Area Code
TAC	Type Allocation Code
TB	Tranposrt Block
TCP	Transmission control protocol

TDD	Time Division Duplex
TDMA	Time Division Multiple Access
TDMA	Time Division Multiplexing
TEK	Traffic Encryption Key
TFT	Traffic Flow Template
TFTP	Trivial File Transfer Protocol
TLV	Type-Length-Value descriptor
TMSI	Temporary Mobile Subscriber Identity
TrE	Trusted Environment
tRS	transparent Relay Station
TTG	Transmitter-Receiver Transition Gap
TTI	Transmission Time Interval
TTR	Time-division Transmit and Receive
TUSC	Tile Usage of Subcarriers
Tx	Transmitter
UCD	Uplink Channel Descriptor
UDP	User Datagram Protocol
UE	User Equipment
UGS	Unsolicited Grant Services
UL	Uplink
UL-MAP	Uplink allocation map
UL-SCH	Uplink Shared Channel
UMTS	Universal Mobile Telecommunications System
UpPTS	Uplink Pilot Time Slot
U-SCH	Uplink Scheduling Channel
UTRA	UMTS Terrestial Access
UTRAN	UMTS Terrestial Access Network
VoIP	Voice over Internet Protocol
WFQ	Weighted Fair Queuing
WG	Working Group
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WRR	Weighted Round Robin

1

Introduction

Without doubt, both cellular phones and the Internet have had a great impact on our lives. Since their introduction in the late 1970s and the early 1980s, the demand for cell phones has had a steady growth in terms of usage and popularity. Initially aimed at "mobilizing" telephony service, mobile communications have gone from bettering voice quality, to adding basic exchanges, to the currently witnessed proliferation of delivering fully fledged multimedia services. This latter evolution was motivated, and made feasible, by the exponential popularity that the Internet has undergone since its introduction to the general public in the mid 1990s. Indeed, the Internet has evolved much since then, and has managed to span the introduction of various multimedia services, ranging from emails and file transfers, to live voice and video streams. By the end of the 1990s, extending Internet services to mobile telecommunications was foreseen as a natural evolution. The many efforts made at the time pursuing such extension – both in the industrial and research sectors can already be seen in today's widely deployed Third Generation (3G) networks. The popularity of today's 3G networks was further strengthened by the introduction of truly smart cellular phones, or smart phones, which featured highly usable interfaces and ease of installation of software applications and packages. Figure 1.1 shows the 3G coverage in the some countries, as calculated by the Organization for Economic Cooperation and Development (OECD) [1].

The advent of a capable mobile Internet has made possible many new services and applications, and has impacted nearly all public and private service sectors. With the recent evolutions of 3G technologies, namely HSPA+, users are able to interact live and through both voice and video with their friends and partners. At the same time, sharing services and social networks resulted in multitudes of text, voice and video statuses and snapshots being constantly uploaded. Users are also able to access their work and financial documents on the go, and connect to their working stations that reside either at their offices or in the Internet cloud, greatly enhancing their productivity over the air. Meanwhile, doctors and caregivers are

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

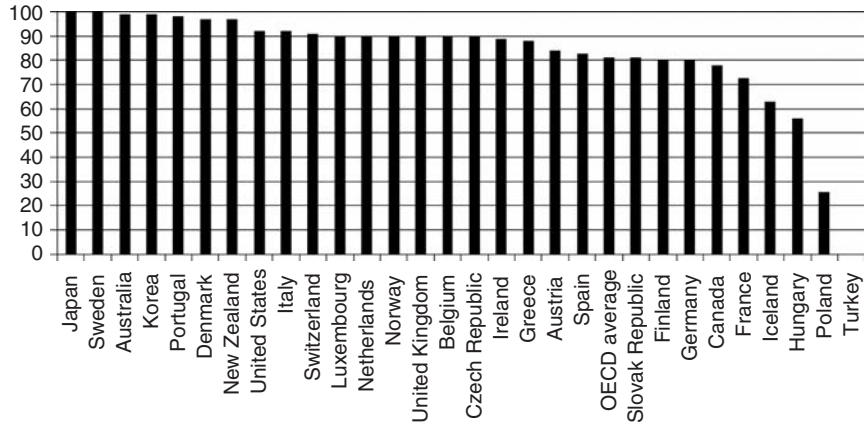


Figure 1.1 3G penetration in various countries up to 2009, per OECD. Note that the average penetration in the surveyed countries is 81 %.

able to monitor the vitals and the state of their patients remotely, immensely reducing costs incurred for commuting and hospital stays costs and improving the patients' overall wellbeing. Third generation networks have also enabled location based services, already being utilized by various targeted advertisements and reward-based credit cards. Such location based services are also enabling the tracking of vehicles, cargo trucks and products nationwide and in real time.

Indeed, much of the above services – and more – can already be witnessed. Despite such possibilities, the increasing demand and popularity of mobile applications and services, in addition to the growing dependence on Internet applications and services in the various sectors (government, commerce, industry, personal, etc.) is calling for a more reliable broadband connectivity that can be made anytime and anywhere. In addition, and as will be noted below, the elemental characteristics of 3G networks hindered their capability of handling this increased demand. Hence, the International Telecommunications Union – Radiocommunications Sector (ITU-R) sought in 2006 to initiate efforts towards realizing more capable networks. The resulting network would mark a substantial improvement over current networks, and facilitate a smooth transition in next generation networks. Such improvements would inevitably include enhancements to both the access network, that is, the Radio Interface Technologies (RITs), and the core network, that is, network management interface.

The intention of this book is to provide an overview of the two Radio Interface Technologies (RITs) that were presented by the Third Generation Partnership Project (3GPP) and the Institute for Electrical and Electronics Engineer (IEEE) in response the ITU-R requirements letter for Fourth Generation (4G), or IMT-Advanced networks. The letter, issued in 2008, identified the target performance criteria in which the candidate technologies must outperform 3G networks. Both candidate technologies, namely 3GPP's Long Term Evolution – Advanced

(LTE-Advanced) and IEEE's 802.16m, were approved by the ITU-R Working Party 5D in October 2010 as initially satisfying the basic requirements.

The objective of this chapter is to elaborate on the motivation for IMT-Advanced networks. The following section summarizes the evolution of the wireless generations, indicating the great advances that have thus far been achieved in wireless communications in general. We next elaborate on the exact motivations for IMT-Advanced. Section 1.3 describes the expected features of IMT-Advanced systems, and the elements of performance used to specify their requirements. Section 1.4 then introduces the two RIT that have been recently approved as satisfying the ITU-R requirements. Finally, Section 1.5 details an overview of the book.

1.1 Evolution of Wireless Networks

Table 1.1. summarizes the history of cellular networks. Through the generations, emphases have been made on different design objectives, ones that best served the requirements of the time.

Interest in the First Generation (1G) cellular, for example, focused on mobilizing landline telephony. The outcome networks, Advanced Mobile Phone Systems (AMPS) and Total Access Communication Systems (TACS), were circuit switched with analog voice transmission over the air. A definite drawback of analog transmission was a generally degraded quality and an extreme sensitivity to basic mobility and medium conditions. Hence, the main design objective in Second Generation (2G) cellular networks was to enhance voice quality. The standards responded by replacing analog voice transmission with digital encoding and transmission, immensely improving voice communication. Improvements to the network core also facilitated the introduction of basic digital messaging services, such as the Short Messaging Service (SMS). The two main

Table 1.1 Generations of cellular technologies [2]

Generation	Year	Network	Technology	Data
1G	Early 1980s	Circuit switched	TACS, AMPS	Analog Voice
2G	Early 1990s	D-AMPS, GSM, CDMA (IS-95)	D-AMPS, GSM, CDMA	Digital Voice
2.5G	1996	Circuit switched or Packet switched	GPRS, EDGE, EVDO, EVDV	Digital Voice + Data
3G	2000	Non-IP Packet switched/Circuit switched	WCDMA, CDMA2000	Digital Voice + High speed Data + video
4G	2012	IP based, Packet switched core network	Not finalized	Digital Voice, High speed Data, Multimedia, Security

standards comprising 2G networks were Global System for Mobile Communications (GSM) and Interim Standard 95 (IS-95), commercially called (cdmaOne). GSM relied mostly on Time Division Multiple Access (TDMA) techniques, while cdmaOne, as the name suggests, utilized Code Division Multiple Access (CDMA). Such division, in addition to variation in the spectrum bands utilized for deployments in different regions, would mark a characteristic interoperability problem that was to be witnessed for a substantial period of time afterwards.

The introduction and the increasing popularity of the 2G technologies coincided with the early years of the Internet. As the Internet experienced an exponential growth in usage, interest in having digital and data services of wireless and mobile devices began to materialize. Evolutions for the two main 2G technologies, GSM into General Packet Radio Services (GPRS) and Enhanced Data Rates for GSM Evolution (EDGE) and cdmaOne into cdmaTwo (IS-95b), enhanced the network cores to be able to handle simple data transfers. For example, GPRS introduced two components, the GPRS Support Node (SGSN) and the Gateway GPRS Support Node (GGSN). The objectives of these components was to augment the existing GSM infrastructure to facilitate data access at the RIT level (SGSN), and to facilitate interconnecting the GPRS network with other data networks, including the Internet (GGSN). Basic email and mobile web access were enabled, but the sophistication of the general mobile Internet experience did not allow popular access, and restricted its usage to the enterprise.

In 1999, the ITU approved five radio interfaces comprising the IMT-2000 technologies. These were the EDGE, cdma2000, Universal Mobile Telecommunication System (UMTS) (Wideband – CDMA (W-CDMA), Time-Division – CDMA (TD-CDMA) and Time Division-Synchronous CDMA (TD-SCDMA)) and Digital Enhanced Cordless Telecommunications (DECT). In 2007, Worldwide Interoperability for Microwave Access (WiMAX) was also recognized as an IMT-2000 technology. These technologies make up the 3G networks. In their design, great emphasis was given to enhance the support for voice services, expand and enhance the support for data services, and enable multimedia to the mobile handset. 3G technologies are sometimes classified based on their nature, with EDGE and CDMA2000 recognized as being evolutionary technologies, that is, enhancing their 2G predecessor technologies, and UMTS and WiMAX as revolutionary, that is, based on completely new radio interfaces. In the case of UMTS, it was WCDMA, while WiMAX relied on Orthogonal Frequency Multiple Access (OFDMA). As will be illustrated in the next chapter, the viability of sub-carrier allocation facilitated by OFDMA has made it the multiple access technique of choice in 4G networks.

3G technologies displayed, and still display, that Internet access through a mobile handset can provide users with a rich experience. The recent widespread of smart phones and pads offered by various vendors indicates the strong demand for such services. However, 3G technologies have faced certain challenges in accommodating the increasing demand. These include deteriorating quality of indoor coverage, unsustainable data rates at different mobility levels, roaming difficulties (incoherent spectrum allocation between different countries), and infrastructure

complexity. While some of these challenges could be efficiently mitigated by denser deployments, the associated cost and complexity made this an unattractive solution. As for network performance, signaling overhead in 3G networks has been observed to consume substantial bandwidths – even more than the requirements of the multimedia being transferred.

The latter revolutions in 3G technology, namely the LTE from 3GPP and the WiMAX 1.5 from the WiMAX Forum, directly addressed these and other issues. Parting away from the RITs that have been used in 2G and early 3G technologies (TDMA and W-CDMA), LTE and WiMAX are based on OFDMA. This facilitated delivering high data rates while being robust to varying mobility levels and channel conditions. The two networks also introduced other technologies, such as using advanced antenna techniques, simplified network core, the usage of intelligent wireless-relay network components, and others.

In early 2008, the ITU-R issued a circular letter initiating the proposal process for candidates for IMT-Advanced technologies. The requirements set for IMT-Advanced were made to address the outstanding issues faced by operators, vendors and users in 3G networks, and were made to accommodate the expanding demand for mobile broadband services. The requirements were set with the general framework of the IMT objectives (i.e., per Recommendation ITU-R M.1645 [3]), which set the desired objectives for users, manufactures, application developers, network operators, content providers, and services providers. Both the 3GPP and IEEE responded with candidate proposals in October 2009, the 3GPP with LTE-Advanced, an evolution of LTE, and the IEEE with the WirelessMAN-Advanced air interface (IEEE 802.16m). Currently, deployments of LTE and WiMAX have already started. The Global mobile Suppliers Association (GSA) indicates commitments by 128 operators in 52 countries [4] in addition to 52 pre-commitments (trial or test) deployments [5]. Meanwhile, the WiMAX forum in its most recent Industry Research Report (IRR) indicates that there are currently 582 WiMAX deployments in 150 countries [6, 7]. Note that these deployments are not IMT-Advanced, that is, are not 4G networks. However, given the ease of upgrade from LTE to LTE-Advanced and from WiMAX 1.5 to WiMAX 2.0, these deployments are indicative of how future deployments will play out.

The initial timeline set by the ITU-R Working Party 5D, the party overseeing IMT-Advanced systems, is shown in Figure 1.2. At the moment, the standardization of both technologies has passed Step 7 which entails the consideration of evaluation results in addition to consensus building and decision. The working party met in October 2010 to decide on the successful candidates and decide on future steps. Both LTE-Advanced and WirelessMAN-Advanced have been recognized as IMT-Advanced technologies. Both standardization bodies are now in Step 8, which entails the development of the radio interface recommendations.

1.2 Why IMT-Advanced

3G networks faced elemental issues in trying to accommodate the projected demand for mobile Internet service. One such issue is the high cost of either

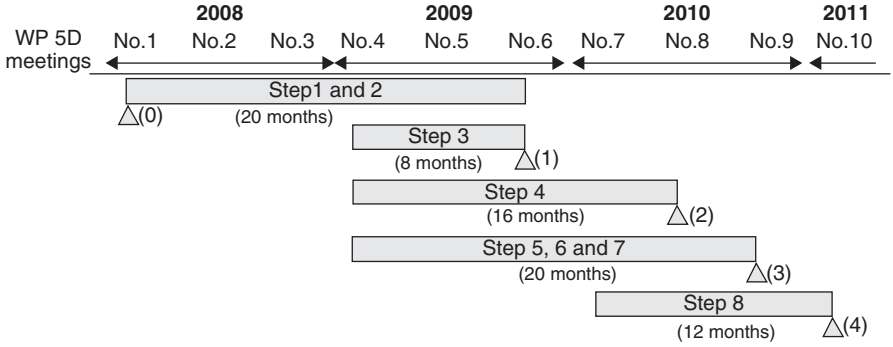


Figure 1.2 IMT-Advanced Timeline.

expanding the network or the network operation in general. Such costs became a substantial consideration when addressing the 3G network performance in densely populated areas or when trying to overcome coverage deadspots. Of particular importance is the performance at the cell-edge, that is, connection quality at overlaps between the coverage areas of neighboring cells, which have been repeatedly remarked to be low in 3G networks. Such problems would usually be addressed by increasing the deployment of Base Stations (BS), which in addition to their high costs entail additional interconnection and frequency optimization challenges.

Certain performance aspects of 3G networks were also expected to be more pronounced. Some aspects were due to the scaling properties of the 3G networks, for example, delay performance due to increased traffic demand. The general support for different levels of mobility also suffered greatly in WCDMA-based networks. Perhaps most critical was the indoors and deadspot performance of 3G networks, especially when various studies have indicated that the bulk of network usage is made while being at either the office or at home.

Combined, the above issues made it cumbersome for operators to respond to the ever increasing demand. Meanwhile, handling specific heterogeneities have made it harder for both operators and user equipment vendors to maintain homogeneous and streamlined service and production structures. For example, the spectrum mismatch between even neighboring countries in 3G deployments prevented users from roaming between different networks – and at times even requiring the user to utilize (and synchronize between) different handsets. At the same time, despite the availability of multi-modal user equipment for a long time, it has thus far been difficult to maintain handovers across the different technologies.

1.3 The ITU-R Requirements for IMT-Advanced Networks

The general requirements for IMT-Advanced surpass the performance levels of 3G networks. Enhanced support of basic services (i.e., conversational, interactive, streaming and background) is expected. Figure 1.3 shows the famous

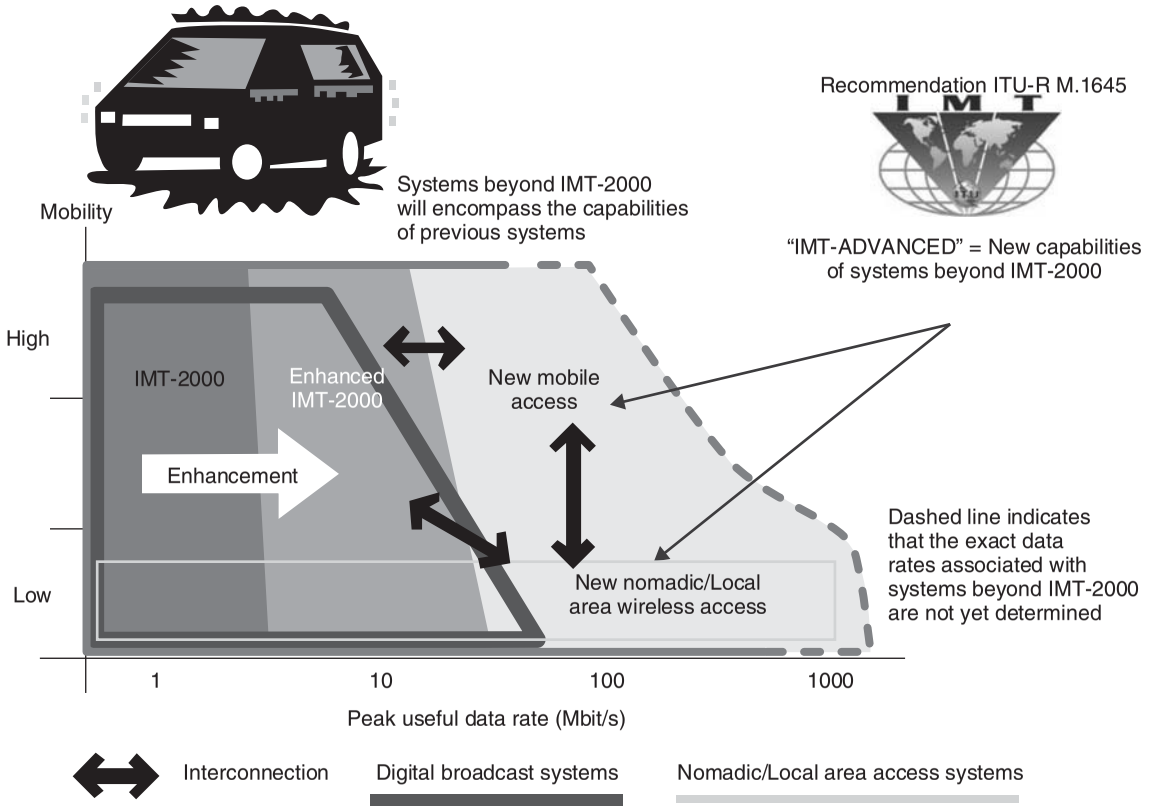


Figure 1.3 3GPP's "van diagram", illustrating the data rates expected to be sustained in systems beyond enhanced IMT-2000 at the different mobility levels.

"Van diagram" which illustrates the relationship between the IMT-Advanced requirements and previous generations. The figure shows the serving area, with one axis being the sustainable data rate supported, while the other shows the mobility level at which that rate can be supported. For example, high mobility (>120 km/h) could only be supported up to ~15Mbit/s in Enhanced IMT-2000. The expectation, per the van diagram, is that the technologies will enable the support of data rates that are at least an order of magnitude higher. For stationary to low mobility (<10 km/h) it is foreseen that data rates surpassing 1Gbit/s can be sustained, while >100 Mbit/s are projected for high mobility levels.

While the data rates are perhaps a key defining characteristics of IMT-Advanced networks, the requirements in general will enable such networks to exhibit other important features, including the following [6].

A high degree of commonality of functionality worldwide with flexibility to support a wide range of services and applications in a cost efficient manner. Emphasis here is on service easiness and application distribution and deployment.

Compatibility of services within IMT and with fixed networks. In other words, IMT-Advanced should fully realize extending broadband Internet activity over wireless and on the move.

Capability of interworking with other radio access systems. An advantage for both operators and users, as it expands the viability of using the RIT most appropriate for a certain location, traffic and mobility. It also strengthens the economic stance of the users.

High quality mobile services. Emphasis here is not just on high data rates, but sustainable high data rates, that is, connection performance that overcomes both mobility and medium challenges.

User equipment suitable for worldwide use. A clear emphasis on eliminating, as much as possible, handset and user equipment incompatibility across the different regions.

User-friendly applications, services and equipment. Ease and clarity of use in both the physical and the virtual interfaces.

Worldwide roaming capability. An emphasis on exploiting harmonized spectrum allocations.

Enhanced peak data rates to support advanced services and applications (100 Mbit/s for high mobility and 1 Gbit/s for low mobility). Such values are to be considered as the minimum supported rates, with high rates encouraged to be sought by the contending candidates.

The ITU-R Report M.2134, entitled "Requirements related to technical performance for IMT-Advanced radio interface(s)" [8], comprises the following elements in specifying the characteristics of future networks.

- Cell spectral efficiency
- Peak spectral efficiency

- Bandwidth
- Cell edge user spectral efficiency
- Latency
  - Control plane latency
  - User plane latency
- Mobility
- Handover Interruption Time
- Voice Over Internet Protocol (VoIP) capacity
- Spectrum

Sustaining a specific rate is more viable at lower speeds than at higher speeds. This is due to the characteristics of the wireless channels and the mobility effects on the quality of the received signal at both sides of the communication link. There are also matters related to sustaining a certain performance level for mobile users during handovers. Accordingly, the IMT-Advanced requirements document indicates the different classes of mobility for which the requirements are defined in order to clarify ITU's expectations. The following classes of mobility are defined.

- Stationary: 0 km/h
- Pedestrian: >0 km/h to 10 km/h
- Vehicular: 10 to 120 km/h
- High speed vehicular: 120 to 350 km/h

The document also identifies the test environments for IMT-Advanced, and the mobility levels supported in each test environment. These are shown in Table 1.2.

It should be noted that most of the values required below are defined assuming antenna configurations of downlink 4×2 and uplink 2×4. For example, a 4×2 arrangement in the downlink means that 4 antennas would be utilized at the base station and two antennas would be utilized at the user equipment or mobile station. Similarly, a 2×4 arrangement in the uplink means that two antennas are utilized for transmission at the user equipment and four antennas at the base station. We elaborate on such advanced antennas setup in Chapter 2.

Table 1.2 Test environments and the supported mobility levels

	Test Environments			
	Indoor	*Microcellular*	*Base coverage urban*	*High speed*
Mobility classes supported	Stationary, pedestrian	Stationary, pedestrian, Vehicular	Stationary, pedestrian, vehicular	Vehicular, High speed vehicular

Table 1.3 The required cell spectral efficiencies in the different environments in IMT-Advanced

Test environment	Downlink (bit/s/Hz/cell)	Uplink (bit/s/Hz/cell)
Indoor	3	2.25
Microcellular	2.6	1.80
Base coverage urban	2.2	1.4
High speed	1.1	0.7

1.3.1 Cell Spectral Efficiency

A cell's spectral efficiency is the aggregate throughput for all users in that cell divided by the nominal channel bandwidth (computed by multiplying the effective bandwidth by the reuse factor), all divided by the number of cells. Table 1.3 shows the requirements values for cell spectral efficiency at different mobility levels.

1.3.2 Peak Spectral Efficiency

The peak spectral efficiency is the highest theoretical data rate (normalized by bandwidth) that can be delivered to a single Mobile Station (MS) when all available radio resources for the corresponding link direction are utilized. The minimum requirements for peak spectral efficiency are 15 bit/s/Hz for downlink and 6.75 bit/s/Hz for uplink. These values are defined assuming antenna configuration of 4×2 for downlink and 2×4 for uplink.

1.3.3 Bandwidth

The candidate technology shall operate with scalable bandwidth allocations using either single or multiple RF carriers, up to and including 40 MHz. Supporting wider bandwidths (e.g., up to 100 MHz) is encouraged by the proponents.

1.3.4 Cell Edge User Spectral Efficiency

The cell edge user spectral efficiency is the average user throughput over a certain period of time, divided by the channel bandwidth. Table 1.4 details the required cell edge user spectral efficiency in the different test environments.

1.3.5 Latency

The requirements specify latencies at both the control plane (C-Plane) and user plane (i.e., transport delay). C-plane latency is defined as the transition time between different connection modes, and is required to be less than 100 ms for

Table 1.4 The required cell edge user spectral efficiency in different test environments in IMT-Advanced

Test environment	Downlink (bit/s/Hz)	Uplink (bit/s/Hz)
Indoor	0.1	0.07
Microcellular	0.075	0.05
Base coverage urban	0.06	0.03
High speed	0.04	0.015

idle to active transition. On the other hand, user plane latency describes the time it takes an IP packet that is ready to be transmitted at one end of the access link (i.e., base station or mobile station) to be ready for processing by the IP layer at the end of the access link (i.e., respectively the mobile station or the base station). The delay latency includes delay introduced by associated protocols and control signaling assuming the user terminal is in the active state. The latency is required to be less than 10 ms in unloaded conditions for small IP packets (e.g., 0 byte payload + IP header) for both downlink and uplink

1.3.6 Rates per Mobility Class

Table 1.5 specifies the expected average spectral efficiencies for mobile users travelling at different speeds. For instance, users traveling at 10km/hr the user can expect a spectral efficiency of 1 (Bits/s/Hz). This translates into a sustained 40 Mb/s given an allocation of 40 MHz.

1.3.7 Handover Interruption Time

Handover interruption time is perhaps one of the most critical requirements in the ITU requirements documents. This is the time in which a mobile handset loses all effective communication (back and forth) as it is in the middle of disassociating with the serving BS and associating with the target BS. Naturally, the duration

Table 1.5 The required rates to be sustained at the different mobility levels in IMT-Advanced

Test environment	Bit/s/Hz	Speed (km/h)
Indoor	1.0	10
Microcellular	0.75	30
Base coverage urban	0.55	120
High speed	0.25	350

has a significant impact on the quality of voice (e.g., VoIP) and video (e.g., video streaming or video conferences) communication. To achieve a requested system throughput increase, the Medium Access Control (MAC) management overhead caused by handovers has to be decreased as well. This reduction can be achieved by changing the MAC management message structure and by changing the MAC message exchange scheme.

For handovers performed while the MS maintains frequency and band – the expected norm for handovers – the interruption time shall not exceed 27.5 ms. If the frequency is not maintained, but the handover is performed with the same band, and an additional 12.5 ms are allowed for frequency assignment. Creating a total bound of 40 ms. If both frequency and band are changed, a 60 ms bound is set.

IMT-Advanced are also expected to support inter-technology handovers to the full extent (both interworking, and across different operators, as applicable). By support, it is expected that sufficient abstractions would be provided by the management functionalities to allow the MS or the operators of the different technologies to facilitate an inter-technology handover. Extending the emphasis to full coexistence; the support should accommodate handovers between IMT-Advanced technologies, in addition to handovers between IMT-Advanced and some selected legacy technologies, namely 2G, IMT-2000, and WiFi. It should be noted, however, that no fixed bounds were made regarding the handover interruption times for inter-technology handovers.

1.3.8 VoIP Capacity

The requirements document defines the VoIP capacity as the minimum of the calculated capacity for either link direction divided by the effective bandwidth in the respective link direction. The values shown in Table 1.6 are derived assuming 12.2 kb/s codec with a 50 % activity factor such that the percentage of users in outage (<98 % packet delivery success within a 50 ms delay bound) is less than 2 %.

Table 1.6 The required voice capacity (in terms of VoIP) calls for different test environments in IMT-Advanced

Test environment	Min VoIP capacity (Active users/sector/MHz)
Indoor	50
Microcellular	40
Base coverage urban	40
High speed	30

1.3.9 Spectrum

A first step in realizing the aforementioned features is to establish, as much as possible, common frequency bands dedicated to IMT and/or IMT-Advanced. The following frequency bands have been recognized by the ITU as ones that can be harmonized across the different regions.

- 450–470 MHz
- 698–960 MHz
- 1710–2025 MHz
- 2110–2200 MHz
- 2300–2400 MHz
- 2500–2690 MHz
- 3400–3600 MHz

1.4 IMT-Advanced Networks

1.4.1 LTE-Advanced

The 3GPP Technical Report (TR) 36.913 [9] details the requirements for LTE-Advanced to satisfy. The document stressed backward compatibility with LTE in targeting IMT-Advanced. It does, however, also indicate that support for non-backward compatible entities will be made if substantial gains can be achieved. Minimizing complexity and cost and enhanced service delivery are strongly emphasized.

The objective of reduced complexity is an involved one, but it includes minimizing system complexity in order to stabilize the system and inter-operability in earlier stages and decreases the cost of terminal and core network elements. For these requirements, the standard will seek to minimize the number of deployment options, abandon redundant mandatory features and reduce the number of necessary test cases. The latter can be a result of reducing the number of states of protocols, minimizing the number of procedures, and offering appropriate parameter range and granularity. Similarly, a low operational complexity of the UE can be achieved through supporting different RIT, minimizing mandatory and optional features and ensuring no redundant operational states.

Enhanced service delivery, with special care to Multimedia Broadcast/Multicast Service (MBMS), will be made. MBMS is aimed at realizing TV broadcast over the cellular infrastructure. It is expected, however, that such services will be undersubscribed in 3G networks. It is hence very critical to enhance MBMS services for 4G networks as it will be a key differentiating and attractive service.

LTE-Advanced will feature several operational features. These include relaying, where different levels of wireless multihop relay will be applied,

and synchronization between various network elements without relying on dedicated synchronization sources. Enabling co-deployment (joint LTE and LTE-Advanced) and co-existence (with other IMT-Advanced technologies) is also to be supported. Facilitating self-organization/healing/optimization will facilitate plug-n-play addition of infrastructure components, especially in the case of relay and in-door BS. The use of femtocells, very short-range coverage BSs, will enhance indoors service delivery. Finally, LTE-Advanced systems will also feature facilitating advanced radio resource management functionalities, with special emphasis on flexibility and opportunism, and advanced antenna techniques, where multiple antennas and multi-cell MIMO techniques will be applied.

LTE-Advanced will support peak data rates of 1 Gbps for the downlink, and a minimum of 100 Mbps for the uplink. The target uplink data rate, however, is 500 Mbps. For latencies, the requirements are 50 ms for idle to connected and 10 ms for dormant to connected. The system will be optimized for 0–190 km/h mobility, and will support up to 500 km/h, depending on operating band. For spectral efficiency, LTE-Advanced requirements generally exceed those of IMT-Advanced, for example, the system targets a peak of 30 bps/Hz for the downlink and 15 bps/Hz for the uplink, while average spectrum efficiency (bps/Hz/cell) are expected to reach 3.7 (4×4 configuration) for the downlink and 2.0 (2×4 configuration) for the uplink. Support for both TDD and FDD, including half duplex FDD, will be made possible. The following spectrum bands are targeted.

- 450–470 MHz
- 698–862 MHz
- 790–862 MHz (*)
- 2300–2400 MHz
- 3400–4200 MHz
- 4400–4990 MHz (*)

The (*) marked bands are not within the requirements of the IMT-Advanced requirements, and some IMT-Advanced may not be supported by LTE-Advanced. These bands are the 1710–2025, 2110–2200 and the 2500–2690 MHz bands.

1.4.2 IEEE 802.16m

As a minimum, the requirements for the IEEE 802.16m [10] entail full support for the IMT-Advanced requirements. This is in addition to backward compatibility with legacy or 802.16-2009 systems. There is also the requirement to enhance service delivery to the mobile users, which involves two objectives. The first is to enhance WiMAX 1.5's Multicast Broadcast Services (MBS), which is similar to 3GPP's MBMS; the second is to utilize Location Based Services (LBS), which are aimed at supporting context-based service delivery. As for the operational features supported IEEE 802.16m, they are similar to those for LTE-Advanced.

Most of the requirements of IEEE 802.16m match those of the IMT-Advanced, including operating in the spectrums set by the ITU-R report. Similar to LTE-Advanced, the IEEE 802.16m is intended to support both duplexing schemes, including half duplex FDD. The standard will also support flexible bandwidth allocations, up to 40 MHz.

1.5 Book Overview

This book is about IMT-Advanced access networks. It begins with two introductory chapters. This chapter provides a brief history and motivation for IMT-Advanced networks, and establishes the requirements for IMT-Advanced networks – as set by the ITU-R. The next chapter, Chapter 2, introduces the physical layer technologies and networking advances that are collectively enabling both IEEE and 3GPP to satisfy the IUT-R requirements in their IMT-Advancements, respectively the IEEE 802.16m amendment and 3GPP'S Release 10. The chapter covers the multi-carrier access technologies utilized in IMT-Advanced networks and their immediate predecessors, including OFDMA and SC-FDMA. It also reviews notions of diversity, adaptive modulation and coding, and frequency reuse, in addition to how wideband transmissions ($<$20 MHz) are made possible using carrier aggregation techniques. Advanced antenna techniques, including MIMO, CoMP, and inter-cell MIMO are also introduced. Finally, the chapter discusses the use of small cells through wireless multihop relaying and femtocells, in addition to access composites will be utilized in IMT-Advanced networks.

The remainder of the book is divided into three parts. The first discusses WiMAX or IEEE 802.16 networks based on the amalgamated IEEE 802.16-2009 documents, which includes the IEEE 802.16j amendment for multihop relay WiMAX networks, in addition to the IEEE 802.16m amendment. The second Part discusses LTE and LTE-Advanced documents based on Release 9 and Release 10 recommendations. The third and last part of the book, "The Road Ahead", offers a multi-faceted comparison of the two technologies, provides a view of the IMT-Advanced market and identifies the future outlook for this next generation cellular networks.

Part I consists of Chapters 3 to 8. Chapter 3 introduces the WiMAX network, its air interface and its network architecture. In doing so, it identifies the differences between the IEEE 802.16-2009 and its m amendment. The chapter also provides a brief overview of the functionalities discussed in the remainder of the part, which is organized as follows.

Chapter 4 describes the WiMAX frame structure, in addition to how addressing and identification are performed. The chapter discusses both the TDD and FDD options, how relay stations are accommodated in WiMAX and new frame structure for IEEE 802.16m better suits the ITU-R requirements. It then discusses how addressing and connections identifications are performed in the two generations. Chapter 5 discusses network entry, connection initialization and ranging. Chapter 6 details WiMAX's quality of service classes, initially defined in IEEE

802.16-2009, in addition to how bandwidth requests, reservations and grants are communicated in the network. Meanwhile, Chapter 7 delves into the details of mobility management in the IEEE 802.16 access networks, including the management between legacy and Advanced WiMax and between WiMAX and other access technologies. Finally, the security aspects of the IEEE technologies are introduced in Chapter 8.

Part II, comprising Chapters 9 to 14, discusses LTE and LTE-Advanced and follows the outline of the first part. In Chapter 9, LTE's air interface and architecture is introduced, including 3GPP's support for femtocells and relay stations LTE-Advanced. The chapter also briefs the reader on the contents of the remainder of the part, which is organized as follows.

Chapter 10 delves into the descriptions of the frame structures utilized in both LTE and LTE-Advanced. It also summarizes how 3GPP network elements are identified. Chapter 11 describes the states and state transition of user equipment, describing the processes for both the idle and connected states, and connection establishment and tear down. As well, the chapter describes the state mapping between access and core signaling. In Chapter 12, quality of service handling and connection management is explained, while Chapter 13 describes intra-network and intra-network mobility management and signaling. Additionally, Chapter 13 gives an overview of LTE-Advanced mobility management for femtocells and relay stations. The last chapter in Part II, Chapter 14, discussed security in 3GPP.

Chapters 15 to 19 make up Part III of book. Chapter 15 offers a comparison between the two standards based on how they satisfy the ITU-R requirements, their functionalities, and their individual use of the enabling technologies described in Chapter 2. Meanwhile, Chapter 16 goes into how each technology attends to the ITU-R coexistence and inter-technology handover requirements. Chapter 17 goes into the quality of service aspects of the IMT-Advanced networks. Specifically, the chapter looks at the two technologies' QoS definitions and handling. A market view of the IMT-Advnaced is provided in Chapter 18, and a future outlook is offered in Chapter 19.

A reader interested in a thorough understanding of the two IMT-Advanced networks and their current and future standing is invited to read all the chapters in their given sequence. Readers interested in any of the individual technologies need only to read the respective part. A head-to-head comparison can be made by reading the relevant chapters, for example, Chapters 10 and 16 for frame structure and network identification; Chapters 13 and 18 for mobility management; and so on. Meanwhile, a reader interested into the comparative analysis of the technologies' current and future status can jump right ahead to Part III.

References

[1] 3G Coverage (up to 2009), available at OECD Broadband Portal, http://www.oecd.org/document/36/0,3746,en_2649_33703_38690102_1_1_1_1,00.html.

[2] HSPA to LTE-Advanced, a whitepaper by RYSAV Research, available at 3G Americas, http://www.3gamericas.org/documents/3G_Americas_RysavyResearch_HSPA-LTE_Advanced_Sept2009.pdf.
[3] Recommendation ITU-R M.1645, "Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000", http://www.itu.int/rec/R-REC-M.1645-0-200306-I/en.
[4] Global LTE Commitments, available at GSA Statistics, http://www.gsacom.com/news/statistics.php4.
[5] LTE Global Map, available at GSA Statistics http://www.gsacom.com/news/statistics.php4.
[6] WiMAX Forum, Industry Research Report, March 2011, http://www.wimaxforum.org/resources/research-archive.
[7] WiMAX Deployments, http://wimaxmaps.org/.
[8] Report ITU-R M.2134, "Requirements related to the technical performance for IMT-Advanced radio interface(s)," http://www.itu.int/pub/R-REP-M.2134-2008/en.
[9] 3GPP Technical Report 36.913, "Requirements for further advancements for Evolved Universal Terrestial Radio Access (E-UTRA) (LTE-Advanced)," http://ftp.3gpp.org/specs/html-info/36913.htm.
[10] IEEE 802.16 Broadband Wireless Access Working Group, "IEEE 802.16m Requirements," http://www.wirelessman.org/tgm/core.html#07_002.

2

Enabling Technologies for IMT-Advanced Networks

In Chapter 1, we discussed the various requirements made by the ITU-R for the IMT-Advanced networks. To realize these requirements, IMT-Advanced technologies need to exploit advances at both the architectural (core) network and the access levels. In this chapter, we review the recent advances at the access level that will be used by the IMT-Advanced networks. These will be referred to when discussing the various functionalities in either IMT-Advanced technologies, that is, LTE-Advanced and WiMAX. However, the treatment given here is far from being comprehensive, and only aims at introducing the fundamentals and the envisioned potentials of these advances.

This chapter is organized as follows. Section 2.1 discusses the fundamentals of multicarrier digital modulation. In particular, it discusses the Orthogonal Frequency Division Multiplexing (OFDM) and two multiple access versions of it, namely Orthogonal Frequency Division Multiple Access (OFDMA) and Single-Carrier Frequency Division Multiple Access (SC-FDMA). Next, Section 2.2 discusses multiuser diversity and scheduling. Adaptive Coding and Modulation (ACM) and how it enables a channel-dependent transmission process are discussed in Section 2.3. Section 2.4 addresses the concept of Frequency Reuse (FR) and its different scenarios. Enabling wideband transmissions using carrier aggregation is studied in Section 2.5, while Section 2.6 focuses on Multiple Input Multiple Output (MIMO) techniques. Relaying and the use of femtocells are studied in Sections 2.7 and 2.8, respectively, while Section 2.9 describes the novel Coordinated Multi-point (CoMP) techniques. The widely used power management techniques are highlighted in Section 2.1, while Section 2.11 concludes by talking about inter-technology handovers.

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

2.1 Multicarrier Modulation and Multiple Access

The requirements for the IMT-Advanced mandate that the utilized multiple access technologies are backward compatible with IMT-2000 (3G) systems. To support different services, both contention and contention-free multiple access should be supported. In addition, as a step towards interference control, FR should be supported. At the same time, in order to accommodate heterogeneity in regulations between different regions, both TDD and FDD duplexing schemes should be supported as well, including half and full duplex FDD.

In order to abide by these requirements while achieving the promised levels of performance, LTE-Advanced as well as WiMAX resort to multicarrier techniques. In particular, three techniques are used, namely OFDM, SC-FDMA, and OFDMA. While WiMAX uses OFDMA in both the uplink and the downlink, LTE-Advanced uses OFDMA for the downlink only while using SC-FDMA uplink.

An advantage of multi-carrier access techniques is their robust communication and stable interference management. In fact, multicarrier techniques facilitate fractional FR which will be discussed in subsequent sections. In addition, they also allows exploiting multiuser diversity at smaller granularities than was ever possible in CDMA-based networks. Another advantage of multicarrier techniques is enhancing system throughput by mitigating the frequency-selective randomness, that is, frequency selective fading. This enhancement is achieved by modulating orthogonal subcarriers, and allows these techniques to support different levels of user mobility and withstand different communication conditions as shall be elaborated further shortly.

2.1.1 OFDM

OFDM is probably one of the most striking advances in access technologies. It facilitates higher transmission rates with a reasonable equalization and detection complexities. This high transmission is achieved through modulating a set of narrowband orthogonal subcarriers. In particular, an OFDM block is built as shown in Figure 2.1. The sequence of L modulated symbols, $x_0, x_1, \ldots, x_{L-1}$, are converted into L parallel streams before taking the N-point Inverse Fast Fourier Transform (IFFT) of each. The possible mismatch between L and N is overcome by zero padding the remaining $N - L$ inputs of the IFFT block. Next, the N outputs, $s_0, s_1, \ldots, s_{N-1}$ are converted back to a serial stream before adding the Cyclic Prefix (CP). Finally, the resulting OFDM block is converted to its analog form prior to sending it over the channel.

Using this architecture, an OFDM block can resist the Inter-Carrier-Interference (ICI) by employing orthogonal subcarriers, that is, as a result of using the IFFT. It is also capable of mitigating the channel time dispersion by inserting the CP. In fact, the insertion of the CP is a widely used technique to create a so called guard period between successive OFDM symbols. The CP is simply a repetition of the last part of the preceding OFDM symbol. The length

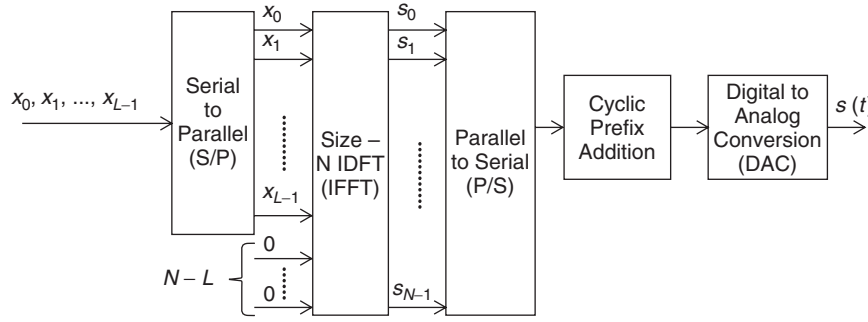


Figure 2.1 OFDM modulation using the IFFT.

of this repetition is made long enough to exceed the channel delay spread, hence mitigating the channel delay spread causing Inter-Symbol-Interference (ISI). In addition, the detection process turns to a circular convolution process which enhances the signal detection capabilities and simplifies the equalization process.

OFDM Demodulation reverses the aforementioned processes. After converting the received signal back into the digital domain, the CP is removed. Next, the signal is converted into a parallel N data streams before performing an N-point FFT. Finally, the sequence is returned back into a serial one. These functionalities are shown in Figure 2.2.

Despite the many advantages of OFDM, actual implementations revealed some challenges. Probably the most famous one is the high Peak to Average Power Ratio (PAPR) problem. Simply put, high PAPR, which results from the coherent addition of the modulated subcarriers, reduces the efficiency of the power amplifier. The high PAPR also sophisticates the Analog to Digital (ADC) and Digital to Analog (DAC) processes [1]. While these two disadvantages can be overcome at the base station side, they form a serious challenge to the battery-powered Mobile Station (MS). Consequently, 3GPP replaced OFDM at the uplink in their IMT-Advanced proposal by SC-FDMA. However, before looking at this novel

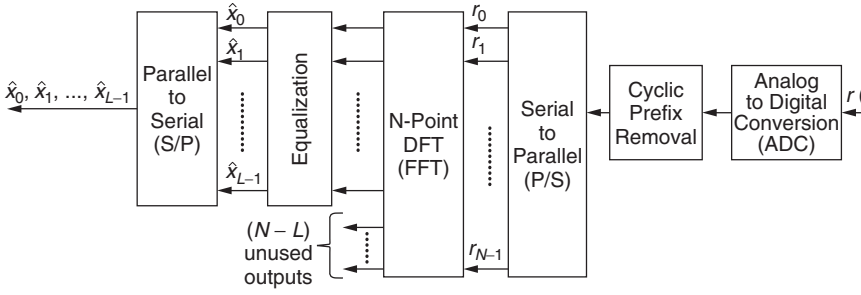


Figure 2.2 OFDM demodulation.

multiple access technique, let us look at the OFDM multiple access version, namely the OFDMA.

2.1.2 OFDMA

In OFDM, all subcarriers are assigned to a single user. Hence, for multiple users to communicate with the BS, the set of subcarriers are assigned to each in a Time Division Multiple Access (TDMA) fashion. Alternatively, an OFDM-based multiple access mechanism, namely the OFDMA, assigns sets of subcarriers to different users. In particular, the total available bandwidth is divided into M sets, each consisting of L subcarriers. Hence, a total of M users can simultaneously communicate with the BS. Subcarrier assignment can be either distributed or localized, as is shown in Figure 2.3.

While in localized assignment, chunks of contiguous subcarriers are allocated to each user, distributed assignment allocates equidistant subcarriers to different users.

Despite the relatively straightforwardness of OFDMA, it has very attractive advantages. Probably the most important of these is its inherent exploitation of frequency and multiuser diversities. Frequency diversity is exploited through randomly distributing the subcarriers of a single user over the entire band, reducing the probability that all the subcarriers of a single user experience deep fades. Such allocation is particularly the case when distributed subcarrier assignment is employed. On the other hand, multiuser diversity is exploited through assigning contiguous sets of subcarriers to users experiencing good channel conditions [2].

Another important advantage of OFDMA is its inherent adaptive bandwidth assignment. Since the transmission bandwidth consists of a large number of orthogonal subcarriers that can be separately turned on and off; wider transmission bandwidths, as high as 100 MHz, can be easily realized.

2.1.3 SC-FDMA

Amongst the many methods proposed and studied to reduce the PAPR of OFDM, SC-FDMA was practically adopted in both, LTE and LTE-Advanced. This OFDM-based multiple access method overcomes the PAPR problem through

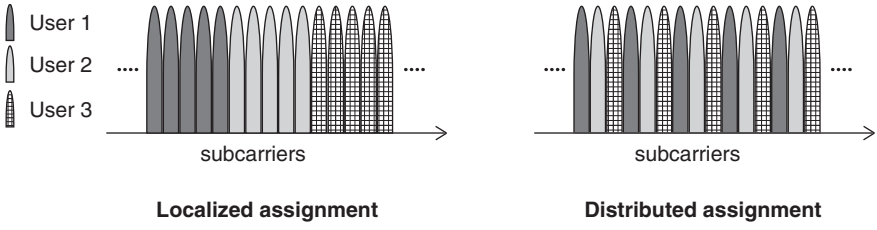


Figure 2.3 Distributed and localized subcarrier assignment strategies.

two additional processes, one at either side of the communication system. More specifically, an L-point Discrete Fourier Transform (DFT) stage is inserted just before the N-point IFFT at the transmitter side, while an L-point Inverse DFT (IDFT) is applied to the L outputs of the N-point FFT at the receiver side.

Since the only modification happens before assigning the different subcarriers, multiple access can be done in a similar way like OFDMA. Accordingly, SC-FDMA possesses the same advantages as OFDMA while experiencing lower PAPR. The adoption of SC-FDMA enhances the power utilization efficiency of the MS batteries, hence prolonging their lifetimes. In fact, LTE-Advanced MSs will use hybrid circuits, where SC-FDMA is used for long-range transmissions, that is, macrocell coverage, while OFDMA is used for short range transmissions, for example, femtocell coverage.

2.2 Multiuser Diversity and Scheduling

In general, wireless channels are prone to random fluctuations caused by the underlying scattering, diffraction, and reflection phenomena. While a passive approach of dealing with these problems would be trying to mitigate their effects, a more fruitful approach involves exploiting these phenomena to enhance system communication. Such exploitation can be achieved by granting access, that is, allocating resources, to the users with good channel quality. Continuous acquisition of Channel State Information (CSI) for all users, however, would be required - a nontrivial process but with substantial gains.

Taking advantage of multiuser diversity in carefully scheduling the users is a way of achieving efficient resource allocation. However, when the scheduling method solely depends on the CSI, it becomes an unfair scheduler since some users may experience bad channel conditions for prolonged periods of time. On the other hand, a scheduling strategy that ignores the CSI of the different users and simply grants them equal access to the shared resources is a fair strategy. However, this fairness is only in the resource allocation and not in the quality of the delivered services. Consequently, a balance should be struck between exploiting the multiuser diversity, fairness in resource allocation, and fairness in service provisioning.

2.3 Adaptive Coding and Modulation

Another dimension where CSI variations are used is the ACM. However, instead of granting or depriving access to the resources based on the CSI, ACM fine-tunes the transmission parameters based on the CSI. Particularly, the coding rate and the modulation order. For instance, if the channel conditions are good, the modulation order is increased while the coding rate is decreased. Similarly, if the channel is bad, a lower modulation order is used along with a higher coding rate. These adaptations give the user the ability to continuously exploit the channel to its best. Hence, achieving a throughput as close to the Shannon capacity as possible.

We should remark here that the benefits of ACM, in addition to those of multiuser diversity in general, do not require perfect (full) or even instantaneous CSI in order to attain the full potentials of the channel. In fact, it has been reported in the literature that using delayed or even incomplete CSI has demonstrated reasonable reliability in both, OFDMA and SC-FDMA systems [3].

2.4 Frequency Reuse

Cellular networks have had always to deal with the interference-capacity tradeoff. Allowing every cell to use all the available spectrum bands boosts the overall system capacity. At the same time, such setup also raises the interference experienced by the cell-edge users to intolerable levels. Consequently, the QoS requirements of these users will not be guaranteed, worsening the overall system fairness. To strike a balance between these two, the notion of Frequency Reuse (FR) is used. In FR, the set of available bands is equally divided between a few neighboring cells, referred to as a cluster. As a result, the system capacity is kept at acceptable levels while the inter-cell interference is significantly reduced.

Nevertheless, full FR, where the entire available spectrum bands are allocated to every cell, remains an attractive option provided that some interference mitigation mechanism is adopted. In fact, full FR can also be applied among the different sectors of the same cell. However, this scenario is prone to higher levels of inter-sector interference especially when the radiation patterns of the different sectors' antennas overlay. Unfortunately, FR cannot be fully realized in practice since inter-cell interference is unavoidable. However, the cell edge problem can be efficiently mitigated through fractional FR. Briefly described, fractional FR divides the available spectrum into a number of segments. Full FR is applied to one of these segments at the cell-center, while the remaining segments are divided between the neighboring cells [4]. For instance, Figure 2.4 illustrates the situation when the entire spectrum is divided into four parts. Observe that while neighboring cells are using different spectrum bands at their edges, the segments used at the center of every cell are the same.

While this scheme enables full FR for part of the available spectrum, the achieved capacity remains below that of full FR over the entire spectrum. This is particularly the case when the segmentation process is static. Alternatively, dynamic segmentation boosts the achievable capacity while providing similar interference mitigation through dynamically allocating the available spectrum resources. Its only drawback is the increased processing complexity at the operator side.

A more recent technique for channel assignment is called dynamic channel assignment (DCA). With DCA there is no fixed association of channels to cells. Each of the channels available to a cell could be used in any sector within the cell as needed. DCA eliminates the need for up-front frequency planning and provides the ultimate flexibility for capacity. However, DCA requires processing and signaling to coordinate channel assignments and avoid interference. Hence, this scheme helps to mitigate interference and improve the network capacity.

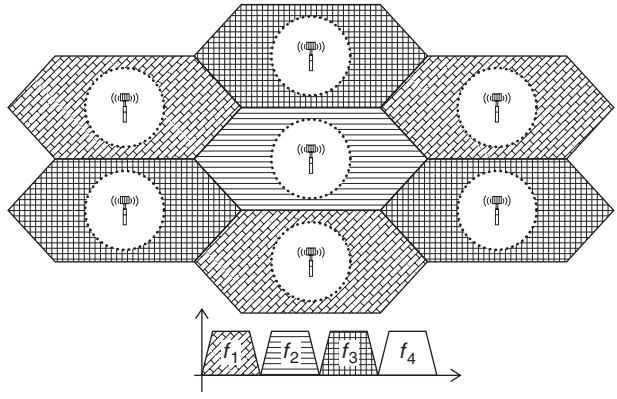
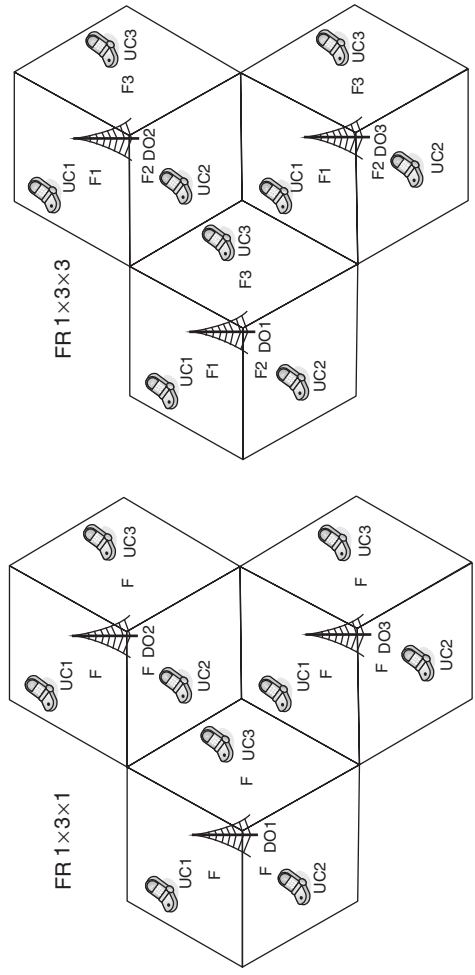


Figure 2.4 Fractional frequency reuse [4].

A large variety of FR schemes can be used in OFDMA system to overcome ICI and improve network performance. These FR schemes are described by the notation $N_c \times N_s \times N_f$, where N_c denotes the number of channels, N_s indicates the number of sectors per BS, and N_f shows the number of fragments in which each channel is divided. When using a $1 \times 3 \times 1$ frequency scheme, there is only one group of channels available to be assigned, and each BS has three sectors. Then, every sector is allowed to use every subchannel in the available frequency as illustrated in Figure 2.5(a). By using this FR scheme, there is no need for frequency planning, therefore simplifying the process for the operator. In$1 \times 3 \times 3$, the available spectrum is divided into three segments: $F1$, $F2$, and $F3$, and each segment is assigned to one sector as shown in Figure 2.5(b). This mechanism simplifies the FR scheme design, because the operator only assigns segments to sectors. Additionally, this FR scheme mitigates ICI by reducing the channel reuse by a factor of 3, however the capacity is also reduced by the same factor.

2.5 Wideband Transmissions

Another area where the flexible spectrum allocation of OFDMA and SC-FDMA systems is exploited is enabling wideband transmission. As has been discussed in the previous chapter, IMT-Advanced networks should support wideband transmissions of as high as 40 MHz, while LTE-Advanced promised supporting even wider transmissions up to 100 MHz. Achieving this while being compatible with 3G networks could be achieved through the so called carrier aggregation. Carrier aggregation refers to the possibility of concatenating several basic (legacy) carrier components into a larger one that can be viewed and managed as a single band. It involves multiple carriers being combined at the PHY layer to provide the user with the necessary bandwidth. The utilization of guard band is possible for the actual data transmission, and utilizing basic (legacy) carrier components achieves backward compatibility with LTE [5].



FR – 1×3×1 (a)

FR – 1×3×3 (b)

Figure 2.5 Sector-based frequency reuse.

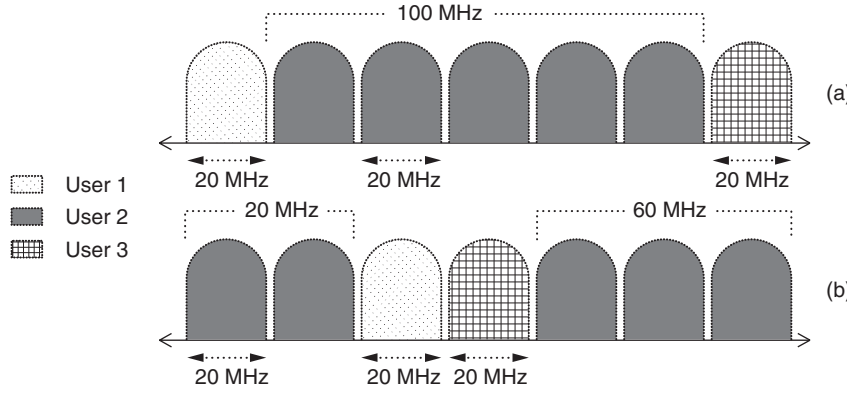


Figure 2.6 Carrier aggregation. (a) Contiguous carrier aggregation and (b) noncontiguous carrier aggregation.

Figure 2.6 illustrates the two possible scenarios for carrier aggregation; namely, contiguous and noncontiguous. The introduction of the latter has been to support carrier aggregation in situations when sufficient contiguous carrier components are not available.

Despite the many promised advantages of carrier aggregation, the associated computational complexities are not marginal. In fact, implementing spectrum aggregation is a challenging process. It requires constant awareness of the available carrier components, the QoS requirements of the different users, the CSI for the requesting users over the different channels, etc. In addition, it requires additional resources to process the aggregation and de-aggregation processes at the PHY layer. For these reasons, spectrum aggregation is expected to be only applied for more capable terminals.

2.6 Multiple Antenna Techniques

Another key technology for increasing the system capacity is the use of multiple transmit and multiple receive antennas, that is, using Multiple Input and Multiple Output (MIMO) techniques [6]. Over the past decade, MIMO techniques became a prominent capacity and reliability enhancement technology for many wireless communication standards including the LTE and WiMAX. Consequently, it is envisioned to remain and even to improve in IMT-Advanced networks.

MIMO techniques involve a variety of techniques aiming at different objectives in different scenarios. In general, they can be divided into Single User MIMO (SU-MIMO) and Multi User MIMO (MU-MIMO), see Figure 2.7. In SU-MIMO, the additional transmit and receive antennas are used to enhance the capacity as well as the reliability experienced by that user. These can be achieved by using space-time codes or beamforming. On the contrary, MU-MIMO generalizes these gains to multiple users. In particular, MU-MIMO exploits the multiuser

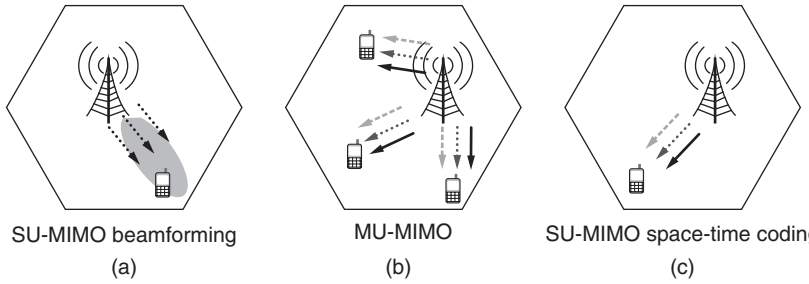


Figure 2.7 Illustration of SU and MU MIMO systems.

diversity in allocating a group of users into the same time-frequency resource [7]. It facilitates achieving high transmission capacity while requiring simpler terminals. To elaborate, data streams on the UL can come from different MSs. The general setting assumes MSs transmitting normally to a BS utilizing more than one antenna. There is generally no coordination assumed between the MSs. Hence, the main challenge is how to schedule the MSs. As MSs are usually dispersed over the cell coverage area and are not predictable in their general behavior, it becomes hard to force a form of control on the non-coordinated behavior while utilizing the possible higher capacities possible.

Another classification of MIMO techniques in cellular systems is the single-site MIMO and cooperative MIMO. While the former encompasses different SU-MIMO techniques, like beamforming, spatial multiplexing, and transmit diversity, and MU-MIMO, the later encompasses the emerging Coordinated Multi-Point (CoMP) transmission and reception [5]. As the names suggest, single-site MIMO is concerned about enhancing the communication experience of in-cell users through employing any of the aforementioned techniques. On the contrary, CoMP aims at improving the communication experience of cell-edge users through inter-cell coordination. CoMP techniques will be described in subsequent sections.

A third classification of MIMO techniques is based on the utilization of CSI at the transmitter side. Under this classification, we have open loop techniques and closed loop techniques. In open loop techniques, the transmitter, the BS or the MS, does not need to have CSI information to adjust its transmission pattern. An example of this category is the single user space-time coding. On the contrary, closed loop systems benefit from the CSI at the transmitter to adjust its transmission parameters. This category encompasses MU-MIMO, SU-MIMO spatial multiplexing, SU-MIMO beamforming, and CoMP. Obviously, exploiting CSI allows the transmitter to achieve transmission rates as close to the capacity as possible at the cost of additional feedback overhead and receiver processing. This is in particular the case when the transmission frequencies of the UL are different from those of the DL.

2.7 Relaying

While a variety of MIMO techniques can be utilized at the BS side, the MS options are limited. In fact, it is physically challenging for a MS to support multiple antennas. Consequently, the network designer needs to find a solution to bring the MIMO virtues to MSs especially at the cell-edge. Such a solution has been found in cooperative communications techniques, also known as cooperative diversity, [8, 9]. Using these techniques, single-antenna MSs can enjoy the MIMO advantages through mutually relaying their signals to the BS. However, this cooperation overcomplicates the processing at the BS side as well as the pricing policies. In addition, it will cause a significant reduction in the MSs battery-lives, which is a critical issue for the users. Consequently, dedicated Relay Stations (RSs) have been proposed to replace the user-cooperation. These stations are not given high-processing or decision-making capabilities; hence they are much cheaper than BSs. As a result, they can be used to increase the coverage area, reduce the transmission range from and to the MSs, hence increasing their achievable throughputs by increasing their Signal to Noise Ratios (SNRs). Unlike BSs, RSs access the network backbone through the BSs. Hence, careful resource allocation strategies are needed. For all these reasons, RSs provide a lower Operational Expenditure (OPEX) and Capital Expenditure (CAPEX) option that allows faster roll out and a flexible configuration.

A variety of RS deployment options can be considered, ranging from the low complexity repeaters to more sophisticated relaying. These variations can help the operator choose the scenario that suits operational needs. The utilization of traditional Amplify-and-Forward (AF) RSs precedes the introduction of IMT-advanced technologies. The role of AF, however, will become increasingly important as it is the most basic and cost-effective form of enhancing communication experience, especially for cell-edge users. AF relays operate in a continuous and nonselective or non-discriminate mode, that is, their operation is not controlled by the BS. On the other hand, Layer 1 (L1) relaying can be selective. Similar to basic repeaters, L1 relays are transparent to MSs. However, the BS controls the RSs transmission power as well as the identity of the RS-served MSs. This also extends to scheduling and retransmission, that is ARQ control.

Both standardization bodies classify RSs based on the deployment objective into two main types, transparent tRSs (Figure 2.8), and non-transparent ntRSs (Figure 2.9). tRS operate within a BS's cell coverage where MSs fully recognize the BS's control message, but have their UL transmission go through the RS. Hence, tRSs aim at expanding the cell's capacity. On the other hand, ntRS are utilized in instances in which MSs are beyond a BS's coverage, and rely fully on the RS for both DL and UL signaling and data transfer. Thus, ntRSs aim at expanding the cell's coverage area.

Note that frequency reuse can also be applied in relay assisted networks, despite being more challenge. The use of a directional antenna at the BS and both

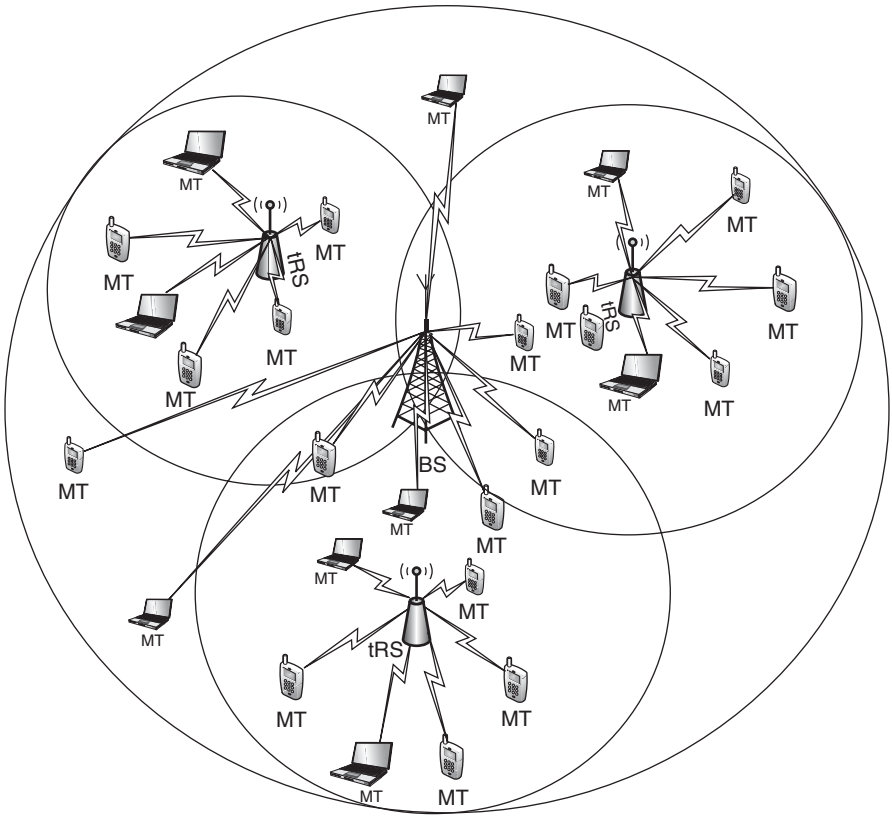


Figure 2.8 Network where transparent RSs are used for capacity enhancement.

omnidirectional and directional antenna at the RS can help to reduce ICI and improve the system capacity. Such a setup would increase the complexity at the BS and RS. Also, in the case of antenna radiation diagrams overlap regions, some frequency channels subsets can be allocated to those MSs in the overlap areas, while other subsets can be assigned to those MSs in the non overlap areas. This mechanism will reduce the ICI in the overlap regions and improve the fairness among MSs in the cell.

2.8 Femtocells

Using RSs, the cellular network could enhance the communication experience of its users by shortening their transmission distances. However, the overall system traffic remained the same since RSs are directly communicating with the BSs. Alternatively, femtocells could reduce this traffic while reducing the transmission range to and from the MSs.

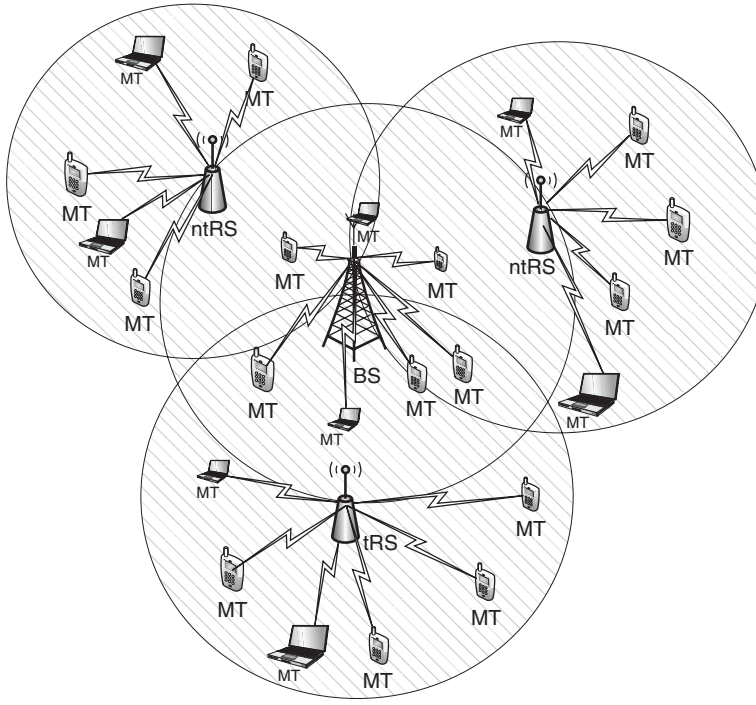


Figure 2.9 Network where non-transparent stations are used for capacity enhancement.

Femtocells have a very little upfront cost to the service provider and can result in very high capacity gains by making use of the wired broadband infrastructure that users already have in their homes/offices. In fact, studies on cellular usage patterns indicate that 70 % of the traffic originates from indoor environments, that is, homes and offices. This traffic requires high SNR levels in order to successfully deliver broadband services, which is challenging because indoor devices operating at gigahertz carrier frequencies suffer from serious attenuation losses. Consequently, the femtocell approach enhances the system capacity with a minimal infrastructure cost. It can also improve the communication quality of the users, hence pleasing both, the operator and the user.

A femtocell can be seen as a small home BS operating in the conventional licensed cellular bands, but with a short-range, a low-cost, and low transmit-power specifications. This customer-instauted BS, shown in Figure 2.10, communicates with the cellular network through a broadband connection (such as DSL or cable) already present at the user's premises. Unlike an RS, a femtocell does not communicate directly with any BS. Since its coverage area is small, a femtocell operates with low transmit power. Consequently, the penetration losses through walls and other infrastructure considerably limit the possible interference to the neighboring BSs and femtocells.

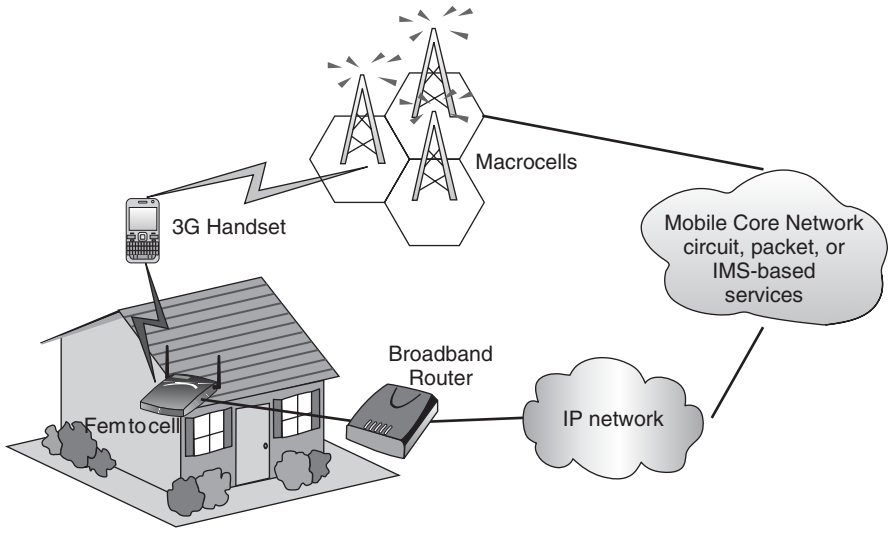


Figure 2.10 An example of a femtocell setup.

Access to the femtocell resources can either be open or closed. A femtocell with an Open Subscriber Group (OSG) allows access to any MS belong to the same network. A femtocell with a Closed Subscriber Group (CSG), on the other hand, allows access only to a limited set of users. A MS would be aware of which femtocells is a subscriber of. Note, however, that all femtocells are required to serve emergency calls, regardless of the MS's subscriber status. Meanwhile, it is possible for CSG femtocell to support a non-subscriber in certain critical instances, for example, priority handovers.

When a pre-authorized MS enters the coverage of a femtocell, it automatically switches affiliation from the serving BS, that is, macrocell, to the femtocell. Hence, initiating as well as receiving calls and data transmissions is performed as usual but through the femtocell instead. Next, this latter encrypts, using the Internet Protocol Security (IPSec), and sends the MS signals through the broadband Internet Protocol (IP) network to one of the switching centers.

Since MSs will be communicating with the local home BS, that is the femtocell, it is envisioned that their battery-lifetimes will be significantly prolonged. Moreover, offloading a fraction of the traffic to the femtocells will improve the macrocell, that is, BSs, reliability since it will be managing lesser traffic. This includes both, service reliability and resource provisioning. Moreover, operation and maintenance costs will also be substantially reduced. While the large operating and maintenance costs of BSs, including site leases, electricity, and backhaul connectivity, can reach as high as $60,000/year/macrocell BS, a femtocells costs as low as $200/year/femtocell. Even when femtocells are deployed in large numbers, it will remain a more economically viable alternative to setting up a new macrocell BS. Finally, a more consistent and satisfactory indoor service is

predicted with femtocells, which will assist both service providers and customers in maintaining long lasting relationships.

Deploying femtocells, however, is not without challenges. A key driver for femtocell unit's cost reduction is the integration of software to silicon, termed "femtocell-on-a-chip," to reduce the number of components needed per femtocell unit. Moreover, operators need to support remote software upgradability to provide service enhancements in a cost effective manner. Femtocell deployment also needs to abide by a single, industry-standard architecture to integrate femtocells to mobile core networks. Currently several femtocell architectures exist for different mobile technologies, for example, WiMAX, CDMA, LTE, which can lead to market fragmentation and hence reduces scalability of femtocell deployment.

Interference management will also be a major concern for femtocells. Femtocells are installed by end-customers through broadband wired connection, which induces lack of coordination, especially frequency planning, between the macrocells BSs and the femtocells. The absence of centralized coordination also causes issues with timing/synchronization between femtocell and macrocell transmissions, which introduces challenges to minimizing interference and carrier offset, and facilitates macrocell-to-femtocell handovers. Meanwhile, macrocell users at the cell-edge might transmit at maximum power, which causes unacceptable interference to nearby femtocells. Similarly, cell-edge macrocell users' DL transmission can be disrupted by nearby femtocells due to their high path loss.

Handover can be a challenge to open-access femtocell deployments, especially macrocell-to-femtocell handover. In open access femtocells, a user might undesirably go through multiple handovers due to channel fluctuations when passing by multiple femtocells, which results in a degraded QoS. Another issue is QoS guarantee over the IP backhaul for delay sensitive traffic. This issue becomes more difficult when a femtocell shares the same connection with Wireless Local Area Network (WLAN), that is, Wi-Fi, traffic. In case of insufficient capacity, a traffic bottleneck where femtocells can experience difficulties transferring data and voice traffic over the broadband connection. Finally, users might choose to relocate their femtocells units outside the femtocell unit's home area. Femtocells operate on the home operator's licensed spectrum, where moving them to a visiting operator's area that uses the same spectrum might raise conflict.

2.9 Coordinated Multi-Point (CoMP) Transmission

The performance of cell-edge users is known to be interference-limited. This has been an unavoidable situation for all previous generations due to the inherent per-cell processing. In other words, since a MS can be only affiliated with one BS at a time, except for the duration of the soft handover process, cell-edge users will be prone to interference from neighboring cells. Consequently, their SINR will be reduced in proportion to the interference level, which directly lowers their achievable capacity. This has been traditionally dealt with by proper resource management between neighboring cells, for example, fractional FR.

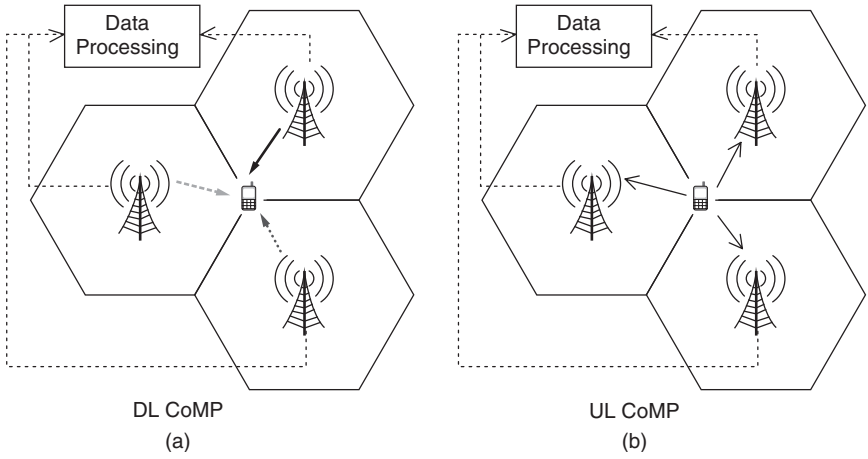


Figure 2.11 DL and UL CoMP techniques.

A radical shift in this cell-centric processing has been recently made through CoMP techniques [11]. CoMP refers to a family of techniques through which the UL and/or the DL transmissions can be simultaneously managed by multiple neighboring BSs as shown in Figure 2.11.

In doing so, CoMP aims at exploiting rather than mitigating inter-cell interference, hence forming a distributed multiple antenna system. By coordinating and combining signals from multiple BSs, CoMP makes it possible for MSs to enjoy consistent performance and QoS when they access and share videos, photos and other high-bandwidth services regardless of their remoteness from the BS. Hence, CoMP techniques improve the cell coverage; enhance the cell-edge throughput, and the overall system efficiency.

As shown in Figure 2.11, CoMP involves multiple, geographically dispersed BSs connected to a central processing unit. This unit is responsible for UL and DL transmission coordination. Despite the very limited literature on these promising techniques, a few CoMP scenarios can be identified.

2.9.1 Interference Cancellation

This is a basic DL CoMP technique. It aims at reducing the amount of interference experienced by cell-edge users through coordinating CSI between the BSs. In particular, if the BSs can share the CSI of their served MSs through the network backhaul, then this information can be used to eliminate (or at least reduce) the inter-cell interference through controlling user scheduling, power allocation and beamforming parameters. This type of CoMP is referred to in [10] as interference coordination. Observe that this basic form of CoMP is still intending to mitigate interference rather than exploiting it.

2.9.2 Single Point Feedback/Single Point Reception

In this type of CoMP, terminals are not aware that transmissions are originating from multiple, geographically separated BSs. The terminal only performs basic measurements and reports it to the serving BS. Based on the channel quality perceived on the user side, the network decides which base station is better suited to transmit to the receiver. For channel estimation, terminal-specific references signals are used. Due to the diversity gain (resulting from the selection of the best candidate transmitter), power is improved. Such a setup would allow for backward compatibility with legacy systems.

2.9.3 Multichannel Feedback/Single Point Reception

The mobile terminal's awareness under multichannel feedback/single point reception is higher than that of single channel feedback. Here, the mobile terminal reports the channel status feedback not only for the serving base station, but also for other bases stations from which it is able to distinguish a DL channel. However, the terminal would still be unaware of the exact processing taking place in the network, and its processing would be the same as that made for single channel feedback/single point reception CoMP.

2.9.4 Multichannel Feedback/Multipoint Reception

The multichannel feedback/multipoint reception setup means a more involved mobile terminal. Here, the terminals are made aware of how the coordination setup, for example, from which base station is it expected to receive a transmission, and can use this information for coordinated multipoint reception. This technique mitigates interference and results in higher SINR values for the network. The disadvantage of this setup, however, is that it requires additional signaling the DL to relay the information for the coordinated transmission.

2.9.5 Inter-Cell MIMO

Probably, the most general form of CoMP is the so called cooperative MIMO or multi-cell MIMO. In this scenario, a group of BSs jointly coordinate the UL and DL communication with a group of MSs. Hence, multi-cell MIMO can be seen as a generalization of the aforementioned MU-MIMO. Depending on the processing efficiency, available capacity, and existing delays, varying levels of MS data can be exchanged between a group of BSs. At one hand, exchanging the CSI of the MSs between the various BSs helps mitigating inter-cell interference. On the other hand, exchanging the MS's traffic between the various BSs allows better reception, reduced or even eliminated inter-cell interference, and better QoS provisioning to all cell-edge users.

2.10 Power Management

Great emphases have been made on power management in the IMT-Advanced requirements in order to reduce the MS power demands. By looking at the above technologies, it can be readily noted that this has been addressed in more than way. For example, replacing OFDMA by SC-0FDMA in the LTE-Advanced UL was made to improve the efficiency of the MS power amplifier. Moreover, the various coverage enhancement technologies, that is, RSs, femtocells, and CoMP, tend to reduce the UL power consumption. In addition, the use of the advanced multiple antenna techniques – through improving the overall interference management – reduces the MSs power expenses when combating rough medium conditions.

However, the standards are also exploiting advances in electronics and RF circuit design through revised definition for MS states in addition to improved scheduling and polling mechanisms. And while the requirements for idle-to-connected duration have been substantially shortened in the requirements for IMT-Advanced networks, they are easily handled by the state-of-the-arty circuit technologies. Furthermore, measures are mode of managing the mobility and cell selection of devices in idle or sleep mode, that is, devices will not require extensive signaling to switch cells while in idle states.

2.11 Inter-Technology Handovers

Advances in supporting IP mobility, coupled with the convergence to an All-IP wireless infrastructure, have made possible the interworking of different access technologies. A key advantage in this viability is the possibility of multi-modal devices to maintain sessions while traversing different access technologies, for example, cellular to WiFi, or LTE-to-WiMAX. Both standardization bodies, in addition to 3GPP2, have defined extensive signaling mechanisms to support both interworking and inter-technology handovers at both the radio interface and the core level. There is also the IEEE 802.21 working group, the purpose of which is to define extensible media access mechanisms that facilitate handovers between IEEE 802 based networks and cellular systems and to optimize handovers between heterogeneous media. It aims at providing link layer intelligence and other related network information to upper layers, or the mobility management entity responsible for handover decision making. The scope of IEEE 802.21 is to cover handover initiation and handover preparation where functionalities include network discovery, selection, and handover negotiation in the former and layer 2 and 3 connectivity in the latter. On the other hand, remaining functionalities such as handover signaling, context transfer, and packet reception, fall into handover execution and thus are out of the scope of 802.21.

References

[1] Ramjee Prasad, *OFDM for Wireless Communications Systems*, 1st ed. Boston, USA: Artech House, Inc., 2004.

[2] Samuel Yang, *OFDMA System Analysis and Design*, 1st ed. Boston, USA: Artech House, Inc., 2010.

[3] P. Bianchi, P. Ciblat, W. Hachem N. Ksairi, "Resource Allocation for Downlink Cellular OFDMA Systems – Part I: Optimal Allocation," *IEEE Transactions on Signal Processing*, vol. 58, no. 2, pp. 720–734, February 2010.

[4] Raymond Kwan and Cyril Leung, "A Survey of Scheduling and Interference Mitigation in LTE," *Journal of Electrical and Computer Engineering*, pp. 1–10, May 2010.

[5] Ian Akyildiz, David Gutierrez-Estevez, and Elias Chavarria Reyes, "The evolution to 4G cellular systems: LTE-Advanced," *Physical Communication*, vol. 3, no. 4, pp. 217–44, December 2010.

[6] Andrea Goldsmith, Syed Ali Jafar, Nihar Jindal, and Sriram Vishwanath, "Capacity Limits of MIMO Channels," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 684–702, June 2003.

[7] Qinghua Li et al., "MIMO Techniques in WiMAX and LTE: A Freature Overview," *IEEE Communications Magazine*, vol. 48, no. 5, pp. 86–92, May 2010.

[8] Andrew Sendonaris, Elza Erkip, and Behnaam Aazhang, "User Cooperation Diversity-Part I: System Description," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1927–38, November 2003.

[9] Andrew Sendonaris, Elza Erkip, and Behnaam Aazhang, "User Cooperation Diversity – Part II: Implementation Aspects and Performance Analysis," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1939–48, November 2003.

[10] David Gesbert et al., "Multi-Cell MIMO Cooperative Networks: A New Look at Interference," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 9, pp. 1380–1408, December 2010.

[11] Ralf Irmer et al., "Coordinated Multipoint: Concepts, Performance, and Field Trial Results," *IEEE Communications Magazine*, vol. 49, no. 2, pp. 102–11, February 2011.

Part One

WiMAX

3

WiMAX Networks

The recent increase in demand for wireless Internet traffic is the result of expanding popularity of applications such as interactive gaming, social networks and TVoIP. This increase is the main drive behind continuous advances in wireless broadband technologies. IEEE 802.16 is the first true technology for fixed, nomadic and mobile wireless broadband access. Since 2001, the IEEE 802.16 working group has been developing new amendments. An effort that was concluded by producing the amalgamated IEEE 802.16-2009 standard in early 2009, and the IEEE's response to the IMT-Advanced requirements and which concluded in March 2011 with the IEEE 802.16m amendment.

The chapter is organized as follows. Section 3.1 introduces the IEEE 802.16-2009 standard, which is matched by both WiMAX 1.0 and 1.5. It provides an overview of the air interfaces described by the standard, in addition to the protocol reference model. Note that the IEEE 802.16j amendment is considered part of the amalgamated IEEE 802.16-2009, and is also introduced in this section. The IMT-Advanced WiMAX, denoted WiMAX 2.0. and based on the IEEE 802.16m amendment, is introduced in Section 3.2, together with the newly defined air interface and the System Reference Model. Section 3.3 provides a detailed overview of Part I of the book, briefing the reader on the frame structure, network entry, quality of service handling, mobility management and security.

3.1 IEEE 802.16-2009

The IEEE 802.16 standard describes several modes of operation, each of which fits a specific deployment objective. In the amalgamated standard document, IEEE 802.16-2009, two modes are described: a mandatory Point-to-Multi-Point (PMP) and an optional Multihop Relay (MR). While both modes describe regular downlink communication, that is, from gateway or base station to mobile

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

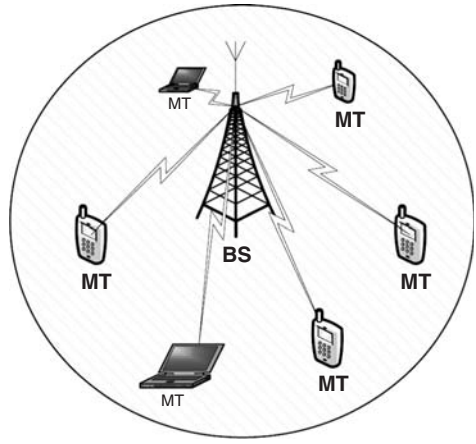


Figure 3.1 A schematic of a IEEE 802.16-2009 deployment, including a base station and different types of mobile terminals.

terminal, the MR mode utilizes intermediate RSs between a cell's BS and the MT. This latter is described in the amendment IEEE 802.16j. An example of an IEEE 802.16-2009 deployment is shown in Figure 3.1.

In a PMP deployment, BSs provide a continuous coverage through a cellular configuration, with the BSs interconnected through a network management infrastructure that oversees the overall management of network operations. Through the BSs, Subscriber Stations (SSs) and Mobile Subscribers (MSs) connect to the network and, when applicable, to the Internet. In the standard, the generic term SS describes user equipment capable of using different RITs operating under both, Line of Sight (LOS) and Non LOS (NLOS) circumstances. On the other hand, MSs are equipment sets whose connected mobility is supported in the NLOS network. As will be described below, mobility is supported only under one IEEE 802.16 interface type, namely OFDMA, and does not require LOS with the BS for communication. More importantly, mobility support is enabled through employing handover mechanisms both within IEEE 802.16 networks and between IEEE 802.16 and other Radio Access Technologies (RAT).

In the IEEE 802.16j amendment, a BS that supports MR is called a MR-BS. In MR, an MR-BS communicates with MSs either directly or through RSs. As was discussed in the previous chapter, a RS is a dedicated, fixed or mobile, relay unit that is connected to the BS through a wireless link. Two types of RS are defined: transparent and non-transparent. Transparent RSs (tRSs) share the carrier frequency with their superordinate station (either an MR-BS or an ntRS) and subordinate stations (only MS), and are mostly deployed within an MR-BS's coverage to improve throughput. Non-transparent RSs (ntRS) are mainly aimed at extending the coverage of an MR-BS cell (MR-cell), and operates either in the same or in a different carrier frequency. When different carrier frequencies are

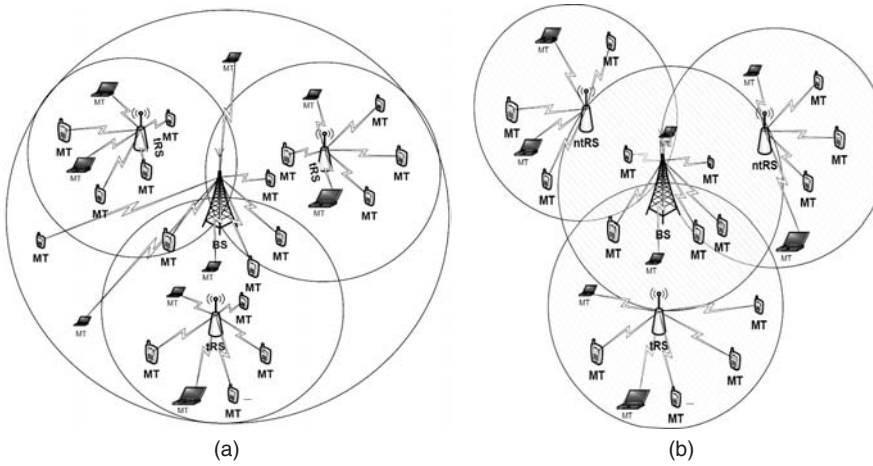


Figure 3.2 Example deployments of IEEE 802.16-2009 relay networks (i.e., amendment j), with (a) showing tRS and (b) showing ntRS.

utilized within an MR-cell or an MR network, the amendment advises the use of interference mitigation mechanisms for both access links (between an MR-BS or RS and an MS) and relay links (between MR-BS and RS or in between RSs). An example MR deployment is shown in Figure 3.2.

Management of the air interface in MR networks can be either centralized or distributed. In centralized operation, all management functionalities are overseen by the MR-BS while in distributed operation, some autonomy is provided for RSs. tRSs always operate in a centralized mode, while ntRSs can operate in both modes. In distributed scheduling, for example, bandwidth allocations for an ntRS's subordinates are made by the ntRS in cooperation with the MR-BS. An autonomous ntRS in distributed scheduling can be also called a scheduling RS.

The IEEE 802.16j amendment is an extension for OFDMA mobility in IEEE 802.16-2009. A salient feature of the IEEE 802.16j is that an MS is not aware of the underlying operating mode of the network, that is, whether PMP or MR. Accordingly, the procedures and signaling made and processed by an MS in both PMP and MR operation are exactly the same. The amendment also describes how MR infrastructure components, that is, MR-BSs and RSs, should handle a MS's requests and traffic in a manner that achieves this seamlessness.

3.1.1 IEEE 802.16-2009 Air Interfaces

The IEEE 802.16 standard describes different air interfaces for different deployment scenarios. For example, the Wireless Metropolitan Area Networks – Single Carrier (WirelessMAN-SC) interface aims at creating wireless backhaul between dedicated stations that rely on LOS connectivity. Meanwhile, the

WirelessMAN-OFDMA, which is our focus in this book, aims at cellular mobile communications.

The following air interfaces are defined in IEEE 802.16-2009:

- WirelessMAN-SC, operates in the 10–66 GHz band with either Time Division Duplex (TDD) or Frequency Division Duplex (FDD) schemes. Moreover, it supports only PMP LOS communications with fixed SSs.
- WirelessMAN-OFDM, operates in the licensed bands below 11 GHz with TDD or FDD duplexing. Supports near-LOS and NLOS communications with fixed SSs but only with provisions for power management, interference mitigation and multiple antennas.
- WirelessMAN-OFDMA, operates in licensed bands below 11 GHz with TDD or FDD duplexing. It supports both, PMP and MR[1] operation. It also supports near-LOS and NLOS communications with either fixed or mobile SSs. In addition, it requires provisions for power management, interference mitigation and multiple antennas.
- WirelessHUMAN, operates in license-exempt bands below 11 GHz (primarily 5–6 GHz) with TDD duplexing. Complies with either the OFDM or OFDMA description. Supports coexistence mechanisms such dynamic frequency selection.

In this book, all descriptions are mainly aimed at OFDM and OFDMA operations. However, exclusive considerations for SC operation will be noted where applicable. No descriptions for WirelessHUMAN will be provided.

3.1.2 Protocol Reference Model

Figure 3.3 shows the protocol reference model for IEEE 802.16-2009. The scope of the IEEE 802.16 standard comprises two planes: Data plane and Management/Control plane. In the Data plane, the standard provides descriptions for both the Medium Access Control (MAC) and the PHY layers. Descriptions for the Management/Control plane include abstractions to be used by Network Control and Management Systems (NCMS). Details of the NCMS are beyond the standard's scope. The described abstractions, however, include descriptions for Service Access Points (SAPs) for both management and control functionalities.

The MAC layer is divided into three sublayers: a Service Specific Convergence Sublayer, abbreviated CS, a Common Part Sublayer (CPS), and a Security Sublayer. Different CSs provide SAPs for upper layers such as ATM, IPv4, IPv6, etc. It also enables classification and processing of higher Protocol Data Units (PDUs) before admitting them to the IEEE 802.16 network infrastructure.

[1] In page 2 of the IEEE 802.16j-2009 amendment, only TDD is mentioned as a duplexing alternative. This is an error as the body of the amendment describes support for both TDD and FDD – including half duplex FDD.

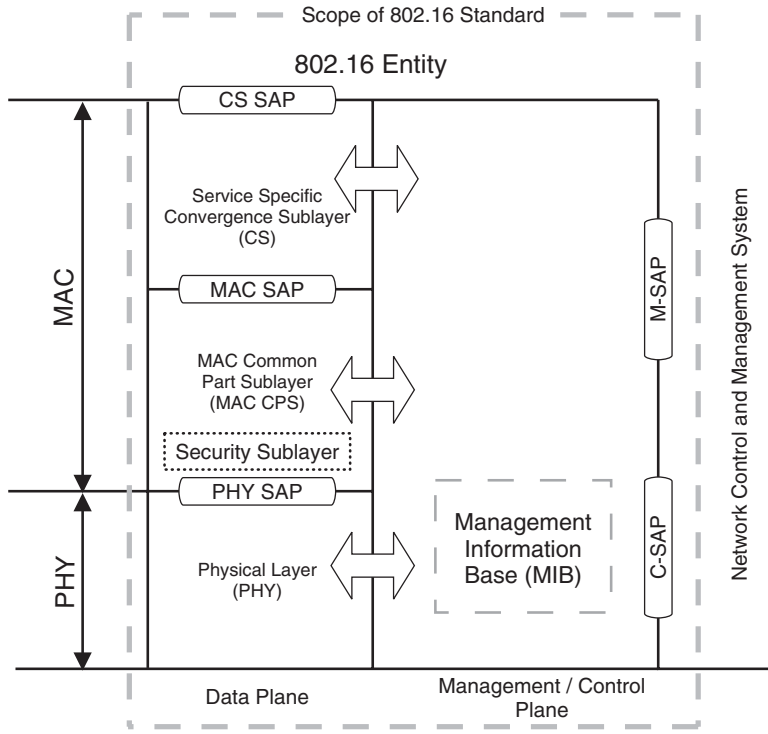


Figure 3.3 The IEEE 802.16-2009 Protocol Reference Model. Reproduced by permission of © 2009 IEEE.

The CPS provides the core MAC functionalities for IEEE 802.16 networks. It receives PDUs from various CSs and applies appropriate classification and Quality of Service (QoS) handling. It also provides a SAP for the different CSs. The CPS also contains a Security Sublayer to provide for communication privacy and integrity.

Descriptions for the PHY layer span the different air interfaces described above. The PHY also offers SAPs for the CPS.

3.2 IEEE 802.16m

The amendment for IEEE 802.16m describes extensive network architecture for Advanced IEEE 802.16 networks. With a refined separation between access and management network services, the IEEE 802.16m provides details for functional entities and interfaces. The IEEE 802.16m network reference model is shown in Figure 3.4.

Similar to the IEEE 802.16-2009, the standard's descriptions is constrained to the access network aspect. The scope of the IEEE 802.16m amendment spans

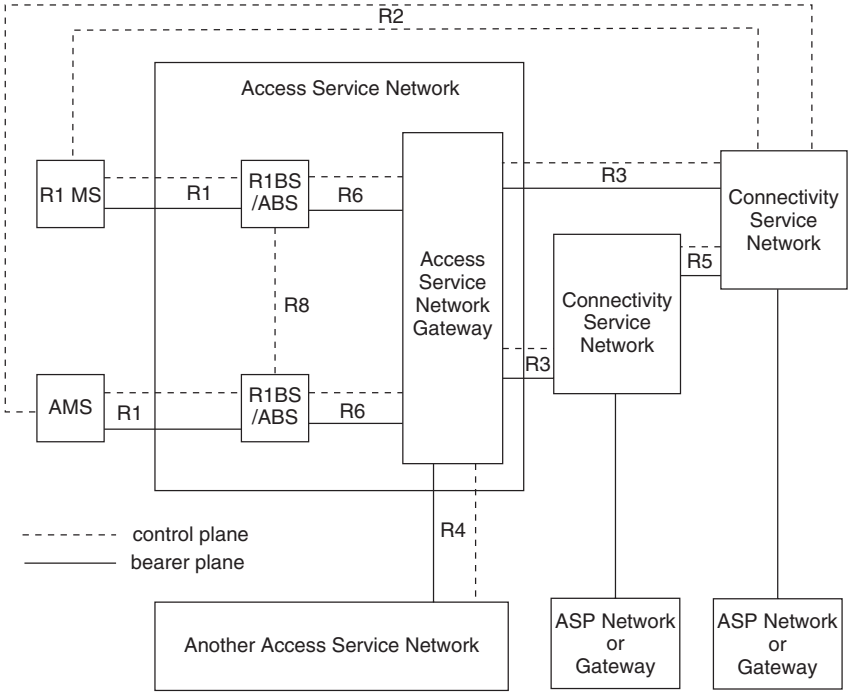


Figure 3.4 The IEEE 802.16m network reference model. Reproduced by permission of © 2009 IEEE.

the description for the Access Service Network (ASN) and the Advanced Mobile Subscriber (AMS). Detailed abstractions are provided for the Connectivity Service Network (CSN), which effectively oversees the higher layer management and connectivity functionalities of the IEEE 802.16m network, in addition to providing the required connectivity for the network's backbone.

Figure 3.5 shows the different deployment examples in 16m. The "purely" advanced network comprises Advanced Base Stations (ABSs) assuming control of the air interface, and through which AMSs (either directly or indirectly) connect to the ASN and CSN. An IEEE 802.16m network may also use Advanced Relay Stations (ARS) as in IEEE 802.16-2009. These can be either, tRSs or ntRSs, based on which centralized or distributed schedulers are used.

An IEEE 802.16m ASN can also deploy Femtocells and support multicarrier operation. An ABS serving a Femtocell is called a Femto ABS. Within our descriptions for IEEE 802.16m, we will interchange the use of Femto ABS and Femtocell. Different types of Femtocells are defined in the IEEE 802.16m amendment. These are differentiated based on which subscribers do they allow access to. A Closed Subscriber Group (CSG) Femtocell is either fully dedicated to its subscriber group (CSG-Closed) or permits users not in its subscriber group but at

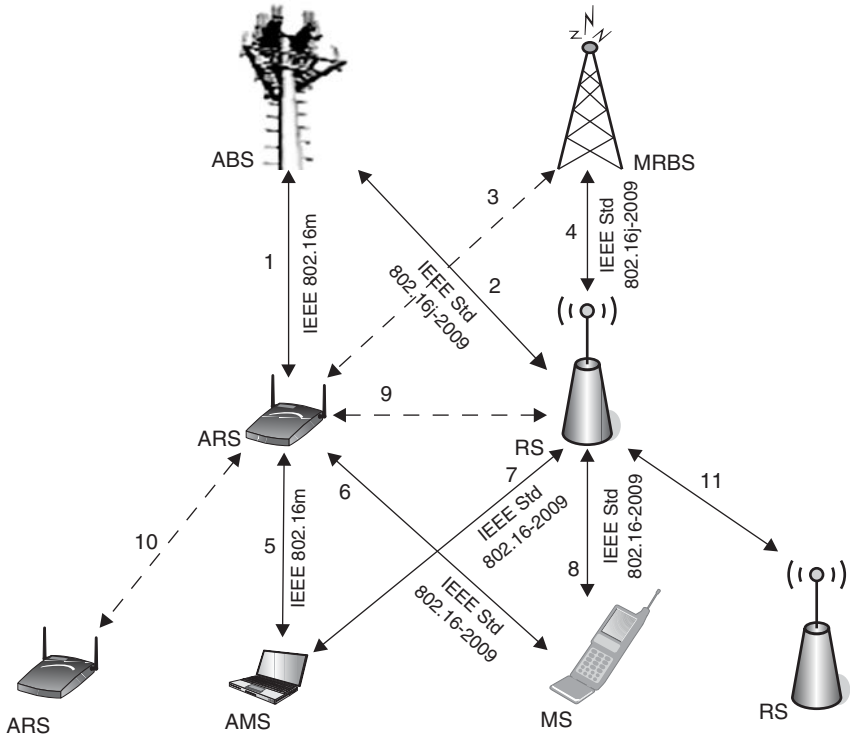


Figure 3.5 An example deployment of an IEEE 802.16m showing different types of links possible between legacy (IEEE 802.16-2009) and advanced IEEE 802.16m) network elements. Reproduced by permission of © 2009 IEEE.

a lower priority (CSG-Open). For the latter, non-SG users will not be attended to if there is a compromise to the Femtocell's resources. On the contrary, an Open SG (OSG) Femtocell is one that is inclusive for all network AMSs belonging to the network. The amendment stipulates mechanisms and conditions for handover between network macrocells and the different types of Femtocells.

A strong feature that IEEE 802.16m also supports is the possibility of multicarrier communication between an access station (Macro or Femto ABSs or ARSs) and an AMS.

Means for backwards compatibility with IEEE 802.16-2009 systems, called Legacy systems by the amendment, are additionally described for the different functionalities. Support for coexistence between Legacy and Advanced systems allows IEEE 802.16-2009 MS, called R1 MS to connect either ABSs or R1 BSs. AMSs that support systems can also enter or handover to an IEEE 802.16-2009 network. Only two connectivity types are not allowed: An ARS connecting through a R1 BS, and mixed (Legacy/Advanced) relay structures.

The IEEE 802.16m also features a strong support for inter-RAT mobility.

3.2.1 IEEE 802.16m Air Interface

The IEEE 802.16m amendment gives a description for a cellular mobile network that satisfies the ITU's requirements for IMT-Advanced systems. The air interface is called Advanced WirelessMAN-OFDMA, and provides support for both TDD and FDD duplexing schemes, including the Half FDD (H-FDD) duplexing.

IEEE 802.16m supports carrier (contiguous subcarriers) and spectrum (noncontiguous subcarriers) aggregations to enable wide transmission bandwidth up to 100 MHz, as per the IMT-Advanced suggestions (The requirement is 40 MHz). Consequently, the channels in IEEE 802.16m do not need to have the same bandwidth nor do they need to be in the same frequency band. These two aggregation processes results in significantly higher peak and average spectral efficiencies than what is achievable by the IEEE 802.16-2009. This capability of IEEE 802.16m, however, entails changes to BS and MS devices and results in high device complexity and challenging resource management.

Another advanced feature of IEEE 802.16m is using extended and improved MIMO techniques. IEEE 802.16m extended the support of multiuser MIMO to eight layers on the downlink and four layers on the uplink. Moreover, it adopted single-user as well as network or multi-cell MIMO techniques. In fact, this latter technique enhances the cell-edge capacity through its inter-cell interference mitigation capabilities. Due to wider bandwidth enhancements included in IEEE 802.16m, the peak spectral efficiency is increased to 17 kb/Hz/s downlink for 4×4 layers and 9.3 kb/Hz/s UPLINK for 2×4 layers.

3.2.2 System Reference Model

The system reference model described by the IEEE 802.16m amendment resembles that of the IEEE 802.16-2009. As can be seen in Figure 3.6, IEEE 802.16m introduces the notion of soft classification whereby a SAP is not required between any two arbitrary classes of functions of the MAC CPS into radio resource control, and between management functions and the MAC. Also similar to the IEEE 802.16-2009 is the categorization of MAC and PHY functionalities into three planes: Data, Control and Management.

3.3 Summary of Functionalities

This part focuses on the elements and functionalities in IEEE 802.16, both Legacy and Advanced, that dictate access network's operation. In what follows, we provide a detailed description of the different chapters and highlight – where applicable – certain features that characterize the IEEE 802.16 and its evolution.

3.3.1 Frame Structure

For both the Legacy and the Advanced IEEE 802.16, the details are provided for an OFDMA frame structure in both duplexing modes, that is, TDD and FDD.

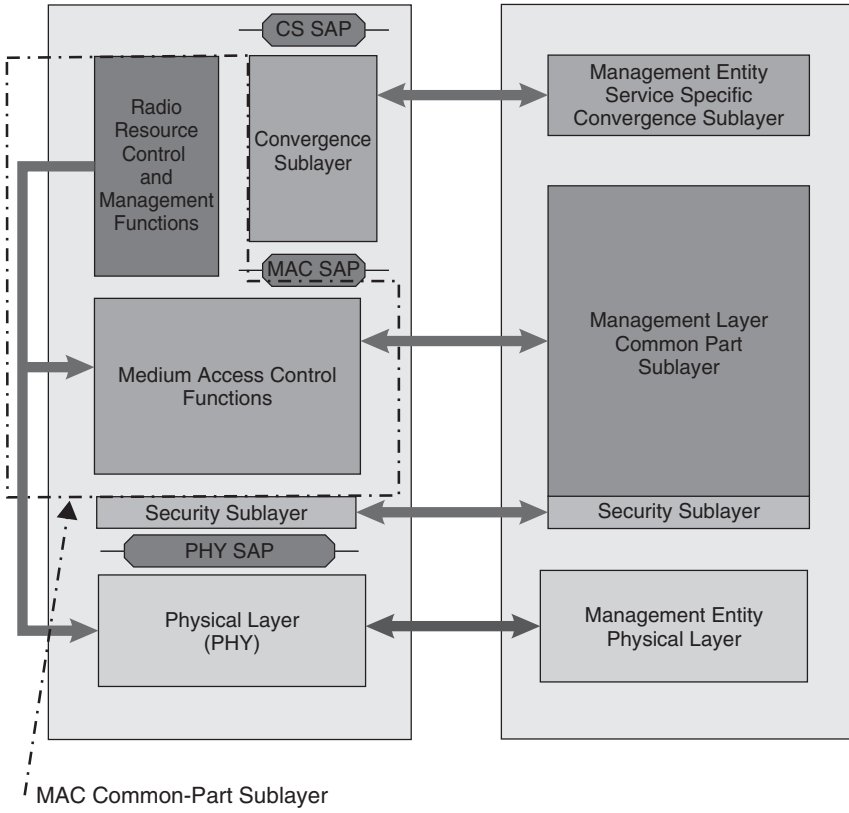


Figure 3.6 The IEEE 802.16m System Reference Model. Reproduced by permission of © 2009 IEEE.

Descriptions for the frame structure operating in both the PMP and the MR modes are also provided. These details and descriptions are discussed in Chapter 4.

The amalgamated IEEE 802.16-2009 document maintains the IEEE 802.16e descriptions[2], allowing for different frame durations. In IEEE 802.16, the BS is the network entity generating the frame, and the one that assumes control of the frame's content for PMP operation. In the MR case, some control can be delegated to the RS, specifically, the ntRS mode where a RS can generate its own frame playing the role of a BS for all MSs associated with it. However, a tRS is always controlled by its superordinate BS. In this case the BS generates an access area and relay area in a frame.

To ensure backward compatibility with the IEEE 802.16-2009 frame structure, the IEEE 802.16m amendment describes a frame structure that achieves legacy

[2] A major difference between the IEEE 802.16e and the amalgamated IEEE 802.16-2009 was the absence of the of the optional "mesh" mode in the latter, clearly indicating that this mode is no longer supported by the IEEE 802.16 WG.

support for IEEE 802.16 deployments, and enables the use of new PHY layer and MAC layer features. The amendment defines a 20 ms length superframe instead of a frame. A superframe is divided into four equally sized 5 ms frames to provide the low latency feature of IEEE 802.16m. Meanwhile, the superframe maintains the support for both TDD and (H-)FDD duplexing modes.

3.3.2 Network Entry

The detailed procedures for network entry, initialization and ranging for IEEE 802.16 are described in Chapter 5.

While network entry, initialization and ranging have distinct operational objectives, their procedures often overlap. Network entry means enabling a SS or MS to associate with an IEEE 802.16 access networks, while initialization involves setting up the elemental connection that would enable the SS to request useful data connections. Finally, ranging is the procedure by which these connections are established. Initial ranging is performed in the user's first entry to the network, while ranging itself is a basic network procedure performed by MSs as well as RSs.

The network entry and initial ranging involve different procedures through which MS connectivity parameters can be configured and primary connections can be established. Downlink synchronization is one of these procedures, during which the MS scans a list of downlink channels to find an active one. The MS synchronizes with a downlink channel by listening to the preamble of the frame transmitted by the BS. The synchronized MS starts of the initial ranging by sending a ranging request to the BS and awaiting a response. The request is sent with a robust modulation and a minimum transmission power. In the event of failure to receive a response, the MS increases the transmission power in an attempt to reach the BS. Once a response is received the MS adjusts its timing and power utilizing the information included in the response message. After a successful initial ranging, the MS informs the BS about its capabilities in a capability request message indicating useful information such as the MCS and the duplexing methods supported. Afterwards, authentication followed by registration takes place and the MS gets connected to the network by assigning it an IP address via the Dynamic Host Configuration Protocol (DHCP) and provide an address for the Trivial File Transfer Protocol (TFTP) server to download any necessary files. The last step is to establish provisioned connections.

Ranging is required at nearly all stages of MS operations to maintain connection quality. Ranging perform mid-connection is distinguished from the initial ranging and is called periodic ranging is the BS and the MS are periodically engaged in its execution.

In the IEEE 802.16m amendment, a concise state machine is defined for both a MS and a RS. This differs greatly from how the MS and RS operations were described up to and including the amalgamated IEEE 802.16-2009 document. In IEEE 802.16m, an AMS transitions between five distinct states; off, initialization, access, connected; and idle.

The network entry and initial ranging procedures are performed while an AMS is in Initialization and Access states. Periodic and other (e.g., for handover) ranging procedures are performed while an AMS is in a Connected state.

In the Initialization state, the AMS performs cell selection by scanning, synchronizing and acquiring the system configuration information before entering Access state. At the Access state, the AMS performs network entry by carrying out multi-steps include ranging, pre-authentication capability negotiation, authentication and authorization, capability exchange and registration. On the success of all of these steps, the AMS receives its Station ID and can now establish initial service flow and transition to Connected state.

When RSs are deployed, they follow similar procedures for connection initialization and maintenance. In addition, RSs may perform interference measurement of neighbor stations, path creation, and tunnel connection establishment with BSs. The RS supports the network entry of an access station by at least providing the initial link adaptation and the remaining network entry procedures may be processed between the MS and the BS in case of tRS.

3.3.3 QoS and Bandwidth Reservation

Quality of serving handling is extensively described in both the IEEE 802.16-2009 and its amendment, IEEE 802.16m. These descriptions are overviewed in Chapter 6 of the book.

In terms of connection-mode, the IEEE 802.16-2009 is designed based on a connection-oriented concept. Each connection is identified by a 16 bit connection ID, with data traffic transmitted over transport connections, while management messages are transmitted over management connections. Transport and management connections are unidirectional and associated to service flows that define the appropriate QoS constraints of the transport connection, and determine the level of treatment the MAC frame receives from the network.

To simplify the design of procedure that determine the level of treatment of MAC frames, IEEE 802.16-2009 specifies five different types of service classes to provide QoS for diverse types of applications.

1. *Unsolicited Grant Service (UGS)*: Supports Constant Bit Rate (CBR) services such as T1/E1 emulation and Voice over Internet Protocol (VoIP) without silence suppression.
2. *Real-Time Polling Services (rtPS)*: Supports variable size real-time data packets generated on periodic basis like Moving Picture Experts Group (MPEG) video or VoIP with silence suppression.
3. *Extended rtPS (ertPS)*: ertPS is a scheduling mechanism that has characteristics similar to both UGS and rtPS. It supports variable-size transport data packets, such as VoIP with silence suppression.
4. *Non Real-Time Polling Services (nrtPS)*: Supports delay tolerant services that require allocations on regular basis, such as File Transfer Protocol (FTP).

5. *Best Effort (BE) Services*: supports BE traffic and provides little or no QoS guarantees such as web surfing over the Internet.

Bandwidth allocations in both, the downlink and the UPLINK, are exclusively managed by the BS. With data to send, the BS schedules the PHY resources required to meet the data QoS requirements on a per-connection basis. Allocations made to specific MS are indicated in the DL-MAP over dedicated management connections. A DL-MAP, as will be detailed later, is a map relayed in the frame to specify the allocations assigned for each MS in terms of both frequency and time slots. Except for UGS connections, a BS may increase or decrease a downlink connection's allocations at its own discretion, that is, to enact certain prioritization policies or to adapt to medium conditions.

For a MS to receive an allocation on the UPLINK, it must generate a Bandwidth Request (BR). A BR may be sent either in a dedicated BR PDU, or optionally piggybacked using the grant management subheader. A BR for a connection can also be incremental or aggregate. For incremental requests, the BS combines the new bandwidth requirements to those of the MS's currently active connection. Aggregate requests, however, are treated as a new view of the MS's total bandwidth requirement. While supporting the incremental BR is optional for SSs and mandatory for BSs, supporting aggregate BRs is mandatory for both. A MS requests bandwidth per connection ID, while the BS processes these requests on a per MS basis.

Naturally, a MS requires allocations in order to be able to send its BR. This process is called polling. BRs for non-UGS scheduling services can be sent using one of the following four methods.

- *Unicast polling*: (applies to SC, OFDM and OFDMA air interfaces) Each MS is individually polled by the BS.
- *Multicast and Broadcast Polling*: (applies to the SC and OFDM air interfaces) Used when there is insufficient resources to individually poll inactive SSs. The BS allocates transmission opportunities and indicates these allocations in the UL-MAP.
- *Contention-based CDMA BR*: (applies to the OFDMA air interface) An MS can use the ranging subchannel and contend using a BR ranging code.
- *Poll Me (PM) bit*: A MS with a currently active UGS connection can indicate that it needs to be polled for non-UGS connections through setting the PM bit in a PDU's Grant Management Subheader GMSH within its UGS connection.

In terms of traffic handling, the standard does not specify any requirements for traffic policing and shaping within an IEEE 802.16 network. As for scheduling, a scheduling algorithm for the UGS service (CBR) traffic, called persistent scheduling, is defined for OFDMA. Persistent Scheduling is a technique used to reduce MAP overhead for connections with periodic and fixed payload-size traffic. UGS resources are persistently allocated by the BS.

The IEEE 802.16m differs from IEEE 802.16-2009 by assigning each flow a four-bit Flow ID (FID). The FID can be combined with a 12-bit Station ID (STID) to generate a network-unique 16-bit identifier for the flow. The objective of introducing the FIDs is to decrease the latency of handover, since the FID does not need to change during handover. Hence, the connections are reestablished faster by just changing the STID from the servicing ABS to the new ABS. However, while IEEE 802.16m is limited to $16(= 2^4)$ connections per MS dictated by the four bits FID; IEEE 802.16-2009 supports up to $65536(= 2^{16})$ connections.

Another distinguishing feature of IEEE 802.16m over IEEE 802.16-2009 is latency reduction by shortening the time needed to honor a BR. The IEEE 802.16m supports an enhanced BR mechanism, where the BR-grant process is reduced into three steps instead of the regular five steps, where step 2 and step 3 are bypassed for faster BR -grant procedure as shown in Figure 3.7.

In addition to the resource management procedures and classifications defined for IEEE 802.16-2009 and IEEE 802.16m, an IEEE 802.16 RS can operate in distributed or centralized scheduling. When a BS is configured to operate in centralized scheduling, the BS schedules all radio resources in its cell. In distributed scheduling, the BS and RS schedule the radio resource on their subordinate links individually. However, the RS produces its schedule given the radio resource assigned to it by the BS.

3.3.4 Mobility Management

A detailed account of IEEE 802.16 mobility management is described in Chapter 7. By definition, mobility management in IEEE 802.16 deals with

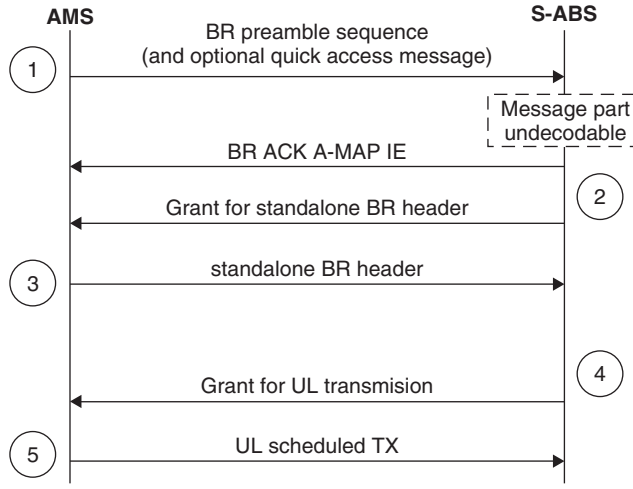


Figure 3.7 The bandwidth request mechanism in IEEE 802.16m. Reproduced by permission of © 2009 IEEE.

realizing seamless mobility as users switch from one part of the network to another, regardless of the utilized access technologies. The standard supports both intra-RAT handovers (Legacy-to-Legacy, Advanced-to-Advanced, legacy-to-Advanced, and Advanced-to-legacy) and inter-RAT handovers (between IEEE 802.16 networks and other access networks such as LTE/-A, HSPA+, WiFi, etc.).

The intra-RAT handover is carried over two phases; network topology acquisition and handover execution phase. Even though the standard does not specify how the handover decision should be made, nor does it mandate whether the decision should be made by the network or the MS, the standard does provide means for information acquisition by both the BS and the MS to make efficient decisions.

The network acquisition phase consists of three steps: network topology advertisement, neighbor BS scanning and association process:

1. *Network topology advertisement* is performed by the serving BS. The serving BS periodically broadcast advertisement messages to all subordinates to provide information about the neighboring BSs.
2. *Neighbor BS scanning*: In this phase the MS acquire the serving BS to specify when and for how long the MS can perform measurements of the neighboring BSs received signal strength. After getting this information, the MS measures the received signal strength of the neighboring BSs after acquiring synchronization with each neighboring BS. After collecting the neighboring BSs measurements and other parameters, the MS decide whether or not a neighboring BS is adequate as a target BS.
3. *Process of association*: This is an optional process, where the MS acquire ranging and service availability information from the neighboring BSs. The objective of this process is to decide on the most proper target BS and to expedite probable future handover.

The second phase is the handover execution phase. It consists of two stages:

1. *Handover preparation*: In this stage, the MS sends handover request message, which includes the measured signal strength, if the neighboring BS received signal exceeds the threshold required for a handover decision. The serving BS communicates to the target BS the expected QoS level for the MS along with the resources required by the MS. The serving BS will choose the best target BS based on the replies received from all target BSs. Subsequently, the serving BS informs the MS by its decision to start the handover action phase.
2. *Handover action*: In this stage, the MS continues the handover by sending a message to the serving BS confirming or cancelling the handover. If the MS decides to disconnect from the serving BS, it stops listening to the serving BS and starts network re-entry with the target BS. Next, it negotiates the basic capabilities, authorization and authentication, registration with the target BS.

Moreover, it terminates its connections' context with the serving BS such as timers, counters, ARQ state-machine, etc. The latency of the handover (which is the duration of the handover action period) can be minimized by if the serving BS sends the MS information to the target BS. Consequently, the target BS can skip some steps of the handover action phase based on the type and the amount of information received from the serving BS.

The IEEE 802.16m supports both, network and MS assisted handover. The handover consists of three phases:

1. *Initialization*: The Initiation phase is only necessary if the handover is being started by an AMS. In this phase, the AMS sends a handover request message to the serving ABS.
2. *Preparation*: The preparation phase starts when the serving ABS sends the AMS information to the target ABSs. This information includes authentication and identification information. Subsequently, the ranging process between the AMS and each target ABS is performed. Based on the information gained by the AMS during the ranging process, the AMS selects a target ABS. Finally, this stage ends when the serving ABS sends control information to the target ABS. The control information indicates whether the handover is hard or soft, the target time of the completion of the handover process and the disassociation with the serving BS, and the AMS connections' information.
3. *Execution*: This phase is similar to the handover action phase of IEEE 802.16-2009. It starts with the network re-entry procedure of the AMS to the target ABS and ends by the AMS disconnection from the serving ABS. If disconnection happens after finalizing the network re-entry, then the handover is soft, otherwise it is hard. Figure 3.8 shows the general handover procedure in an IEEE 802.16m network.

Besides the ABS to ABS handover, IEEE 802.16m defines three other types: R1BS to R1BS (legacy to legacy), ABS to R1BS (IEEE 802.16m to legacy), and R1BS to ABS (legacy to IEEE 802.16m). In addition, it supports handover from and into IEEE 802.16m Femtocells.

IEEE 802.16m defines handover procedures and signaling for handover from IEEE 802.16m network to other RAT (i.e., Inter-RAT handover) including LTE, IEEE 802.11, GSM/EDGE, 3GPP2, UTRA and CDMA-2000. Also, it supports the 802.21 standard for technology independent handover.

The handover in relay networks is not much different from IEEE 802.16-2009 and IEEE 802.16m. In relay network, the A/BS still carry out the scanning and network topology advertisement while the A/RS relays only the MAC control signaling such as the handover command message and indication message between the subordinate AMS and the ABS. If the handover is carried out between the ABS and the ABS's subordinate ARSs, that is, the AMS roams under the ABS, the AMS context information transfer can be omitted.

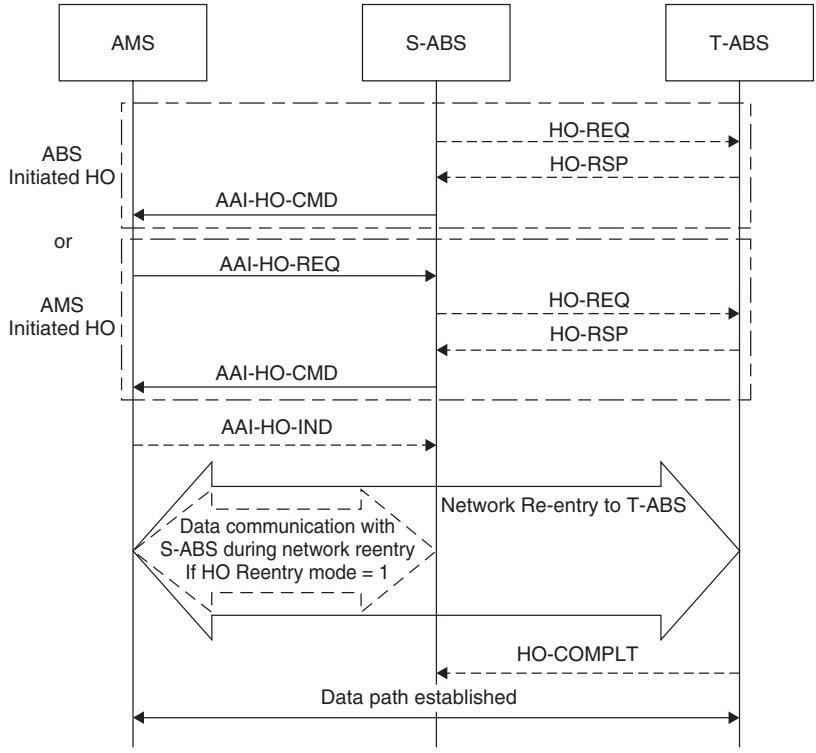


Figure 3.8 The general handover procedure in IEEE 802.16m. Reproduced by permission of © 2009 IEEE.

3.3.5 Security

Chapter 8 describes the robust security functions defined in IEEE 802.16. The functions include strong encryption and mutual authentication. The standard relies on a concept similar to that of the IPSec protocol, known as Security Associations (SA). It defines security parameters such as keys and indicators of the utilized encryption algorithms. It also defines the parameters for unicast services called data SA, multicast services called group SA, and authorization called authorization SA. These latter provides security parameters used in authentication and key establishment necessary to configure the data and the SAs. A SA is established for each service provided by the cell.

An authorization security association comprises of four elements: (1) An X.509 certificate to certify devices in the network; (2) an authorization key for BS/MS authentication; (3) an encryption key, derived from the authorization key, to encrypt traffic during encryption key exchange; (4) a message authentication code, derived from the authentication key and used to authenticate management messages flowing between the BS and the MS.

The data SA provides parameters used for secure data transmission. Data SAs are three types: primary SA, static SA and dynamic SA. A MS has a unique primary SA and zero or more static and dynamic SA. A Primary SA is established between each MS and BS during the initial ranging whereas Static SA is established for each service defined by the BS. Moreover, a dynamic SA is associated with services flows; that is, established and tear down with the establishment and tearing down of service flows.

The parameters include SA identifier to identify each established data SA connection, encryption cipher definition used to provide wireless link confidentiality, traffic encryption key used to encrypt the data messages, and data encryption SA type indicator identifies the data SA type.

Group SAs are used to provide the required parameters to secure multicast traffic. The parameters include group traffic encryption key used for the encryption of the multicast traffic. Group key encryption key used to encrypt the Group traffic encryption key used in multicast traffic.

IEEE 802.16-2009 networks provide security services through three phases: authentication, key establishment and data encryption: Authentication is the process of verifying the identity of devices joining a network. During the authentication phase keying material is exchanged between the MS and the BS, which facilitates the secure exchange of data encryption keys. Data encryption keys are used to ensure the data transmission confidentiality of the IEEE 802.16-2009. IEEE 802.16-2009 does not provide confidentiality protection for the management messages.

The security mechanisms used in IEEE 802.16-2009 are similar to those of relay. To support the multihop functionality, additional security procedures is integrated into the relay standard. The standard defines the concept of a Security Zone. The Security Zone defines security parameters and relations within the relay zone; that is, the BS, RSs and the MSs.

4

Frame Structure, Addressing and Identification

The IEEE 802.16-2009 standard describes the frame structure for the several duplexing modes employed. A BS in IEEE 802.16 assumes control of a frame's contents for PMP operation. In the MR setting, however, it is possible for RSs to assume some control for the areas they cover. This is especially the case when distributed scheduling is used. Moreover, the IEEE 802.16m amendment extends these possibilities for Advanced networks.

This chapter describes the frame structures for the different physical layer technologies in WiMAX, together with the relevant duplexing techniques. It also discusses how addressing and flow identification are performed based on the descriptions given in the standard. The chapter is organized as follows. Section 1 describes the frame structure in the IEEE 802.16-2009 TDD and FDD modes. The frame structure in the IEEE 802.16j is described in Section 2 including both, transparent and non-transparent relaying, while the in IEEE 802.16m frame structure is described in Section 3. Finally, the addressing and connections identification processes are described in Section 4.

4.1 Frame Structure in IEEE 802.16-2009

The smallest unit of resources in IEEE 802.16 is the slot, which can be allocated to a single user. Each slot consist of one subchannel utilizing one, two, or three OFDM symbols, depending on the type of subchannelization used. The standard defines the following four subchannelization types:

- *DL Full Usage of SubCarriers (FUSC)*: Each slot constitutes one subchannel by one OFDM symbol. A single subchannel consists of 48 subcarriers that are not necessarily contiguous;

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

- *DL Partial Usage of Subcarriers (PUSC)*: Each slot constitutes one subchannel by two OFDM symbols. One subchannel consists of 24 data subcarriers grouped in two clusters;
- *UL PUSC* and *Tile Usage of Subcarriers (TUSC)*: Each slot constitutes 16 subcarriers by three OFDM symbols; and
- *Band AMC*: UL and DL contiguous subcarrier permutation. Each slot consists of either eight, 16 or 24 subcarriers by respectively six, three, or two OFDM symbols.

A frame comprises slots in both time and frequency. In the frequency domain, a frame is divided into segments; while in the time domain 6, it is divided into zones. Different frame durations (2.5, 4, 5, 8, 10, 12.5, and 20 ms) are supported; each consisting of a fixed number of slots. Once the network operates with a particular frame duration, it should not be changed as otherwise resynchronization would be required between all network elements. Adjacent slot groups are assigned to users based on their overall demand, individual QoS and traffic requirements, and individually perceived channel conditions. Slots assigned to one user are called the user's data region.

4.1.1 TDD Frame Structure

The frame structure for the TDD operation is shown in Figure 4.1. The frame preamble and the Frame Control Header (FCH) precede the DL-MAP and the UL-MAP. The preamble is positioned at the start of the MAC frame because it is used by the SS physical layer in some operations such as frequency synchronization, time synchronization, and channel equalization. The FCH carries the configuration messages such as the length of the UL-MAP and the DL-MAP, the modulation and coding used, and the available subcarriers.

DL data bursts follow the UL-MAP and DL-MAP messages, and are broadcast to all SSs using Time Division Multiplexing (TDM). Data transmitted to each SS is modulated and coded based on the channel quality between the BS and each SS, and are sorted based on transmission robustness. This means that DL bursts with the most robust modulation, that is, BPSK, are transmitted first. The FCH is modulated with QPSK for robustness, while the MAP is modulated with BPSK and ($^1/_2$) coding rate for reliability. Since modulating the complete MAP in BPSK with (1/2) coding rate can result in a high overhead, the standard provides an option where AMC is used for MAPs aimed at certain SSs, while MAPs intended for all users can be compressed.

The UL subframe starts with a contention region for SSs to perform initial ranging and send their bandwidth requests. The BS specifies an UL interval during which bandwidth requests can be made. Collisions are possible in between SSs and MSs, and are resolved through a random backoff time. SS/MS transmission opportunities or grants follow the contention region and are used to transmit SS bursts. The time slot allocation for each SS is broadcasted to all SSs by the BS in the UL-MAP. To allow sufficient time for radio switching, profile bursts

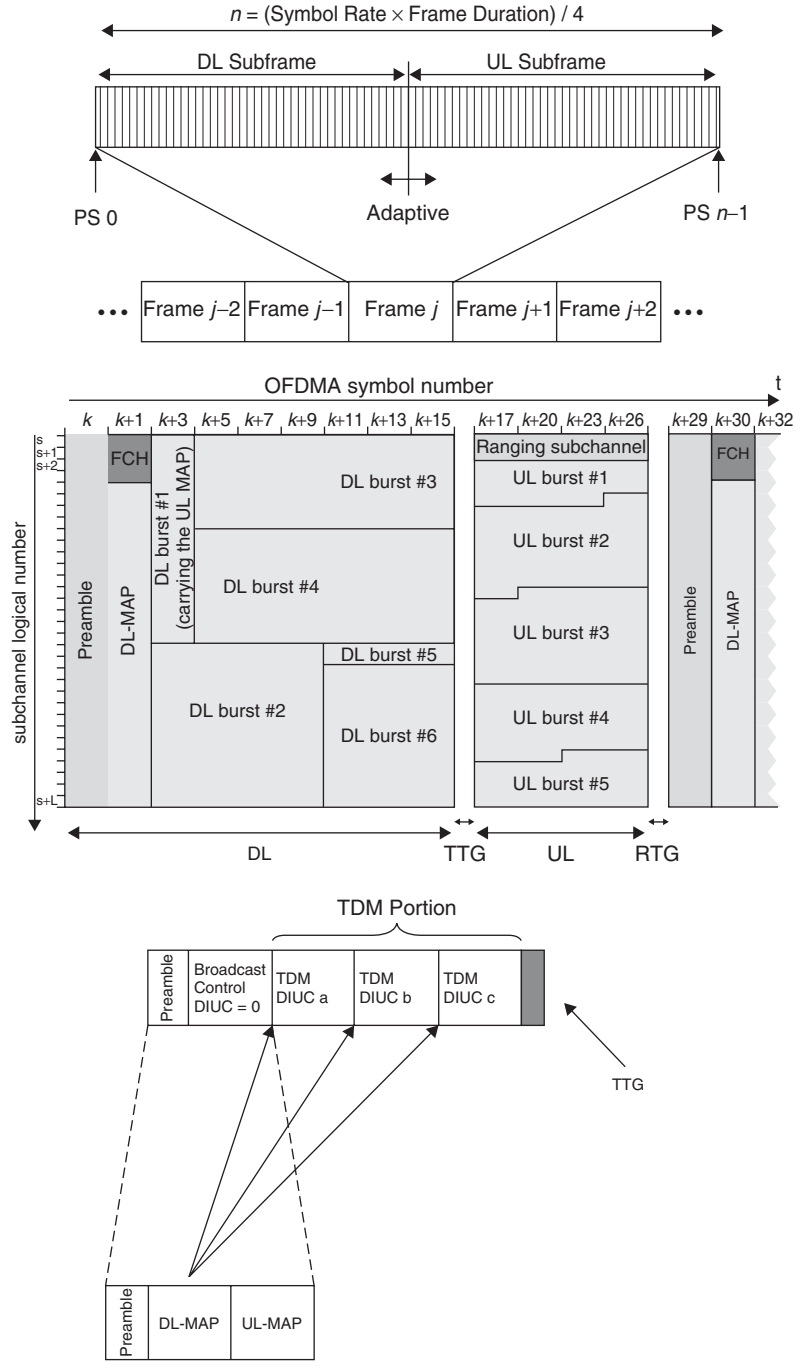


Figure 4.1 OFDMA and OFDM frame structure, TDD operation mode. Reproduced by permission of © 2009 IEEE.

for neighboring SSs are separated by a Transmitter-Receiver (Tx-Rx) Transition or Turnaround Gap (TTG) or SSTTG. An SSTTGG starts with a preamble that separates the different SSs transmissions for BS-SS synchronization.

4.1.2 FDD/HD-FDD Frame Structure

In FDD, the UL and DL subframes are simultaneously transmitted over different carrier frequencies. This is known as full FDD. Half FDD, however, occurs when the transmission and reception times are different. Accordingly, sufficient time needs to be allowed for an SS to switch between the transmitter and the receiver modes. The full and half FDD frames are shown Figure 4.2 below. Note that, apart from the UL and DL being exchanged over two different carrier frequencies, the content of the FDD frame is similar to that of the TDD.

4.2 Frame Structure in IEEE 802.16j

The frame structure in IEEE 802.16 relay mode bears many similarities to that of the PMP. A relay frame is divided into UL and DL parts, it can be TDD based, full FDD based, or half FDD based. However, the UL and DL segments in IEEE 802.16j are divided into multiple time zones depending on the type of communication required, that is, either access or relay. Access zones are dedicated to communications between a MS and either, a BS or a RS, while relay zones are dedicated to communications between a RS and either, a BS or another RS.

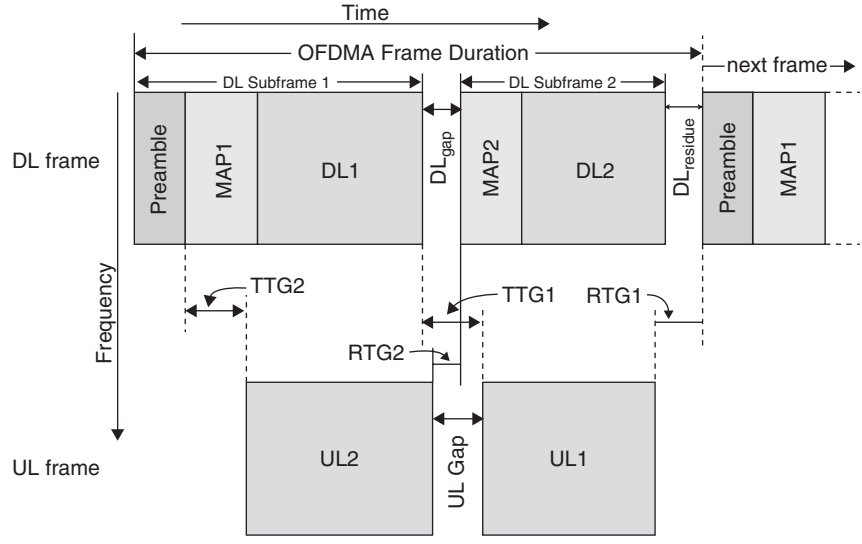


Figure 4.2 OFDMA and OFDM frame structure FDD operation mode. Reproduced by permission of © 2009 IEEE.

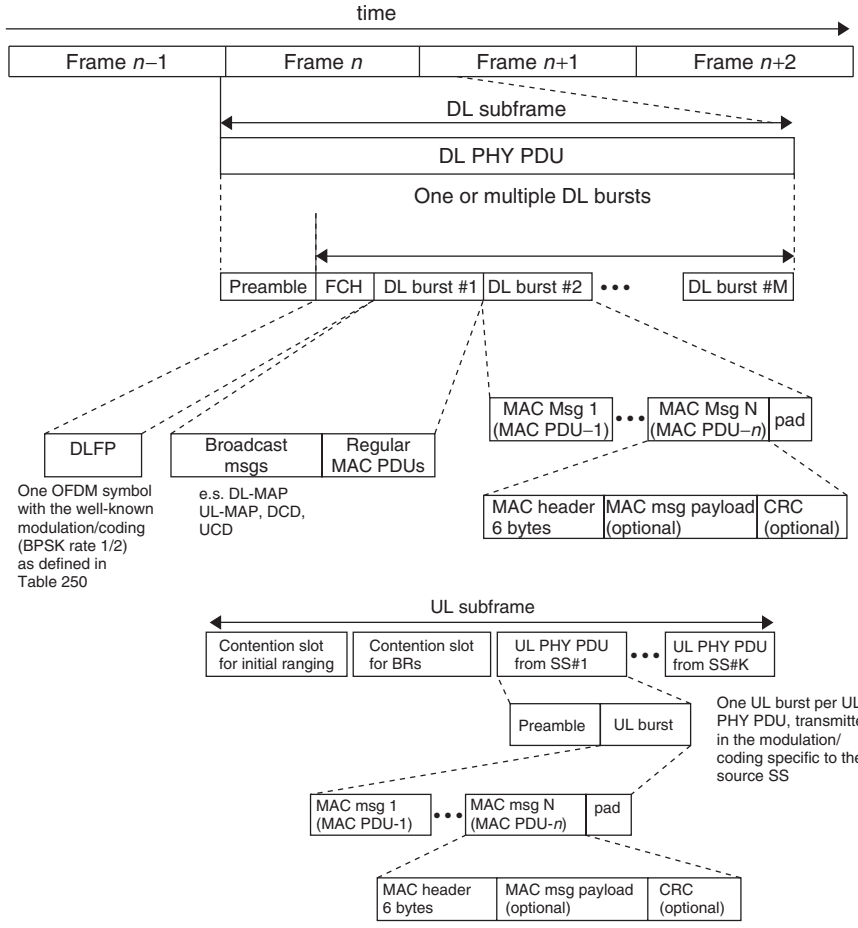


Figure 4.2 (*continued*)

As mentioned earlier, relaying can be either transparent or non-transparent. Only two hops are allowed in the transparent mode (using a tRS), that is, BS-tRS-MS, with MS being within the coverage of the serving BS. On the other hand non-transparent relaying is used for coverage extension and can support two or more hops, that is, ntRSs can serve MSs as well as other RSs (tRS or ntRS). Accordingly, the frame structures of these two modes are expected to be different.

4.2.1 Frame Structure in Transparent Relaying

Figure 4.3 shows the OFDMA frame structure in the transparent relaying case. As in the PMP operation, a TTG separates the DL from the UL in the TDD

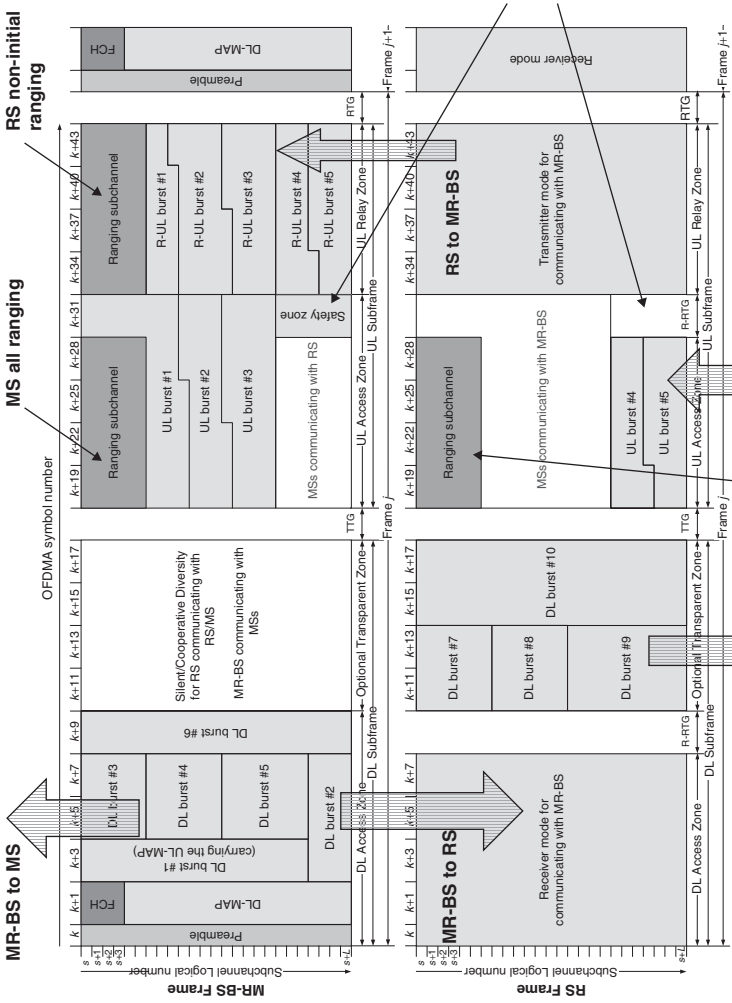


Figure 4.3 Example of a frame structure for transparent relay © IEEE 802.16j-2009, Part 16: Air Interface for Broadband Wireless Access Systems Amendment 1: Multiple Relay Specification. Reproduced by permission of © 2009 IEEE.

frame. Any two consecutive frames are also separated by a Receiver-Transmitter Transition Gap (RTG). The transparent relay frame is different, however, as it is divided into four subframes: a DL access zone, a transparent zone, a UL access zone and a UL relay zone. Having a transparent zone is optional, and is aimed at transmission from the tRS to the MS. The safety zone and the Relay RTG (R-RTG) are used to allow the RS to switch from receiving to transmitting. If the transparent zone is used to forward DL data bursts from the BS to the MS through the tRS, the data may be transmitted from the BS to the RS in one frame. At the same time, the same data can be forwarded from the RS to the MS in a subsequent frame. In this case, a Relay MAP (R-MAP) may be used. Note that, for transparent relaying, the use of R-MAPs is optional.

The DL access zone is populated by control information including the preamble, FCH, UL-MAP, DL-MAP, R-MAP and the Downlink and Uplink Channel Descriptors (DCD/UCD), all of which are directly received by MSs and RSs within the BS's coverage area. The UL/DL-MAPs contain detailed information about the radio allocation for a RS in the access zone, while the R-MAP indicates allocations in the relay zone. A MS transmits data bursts to the BS through the tRS in the UL relay zone, which allows using modulation techniques with higher data rates. The tRSs then retransmits the received data from the MS within its coverage to the BS in the relay zone. In a relay zone, a BS either remains silent or is involved in a cooperative transmission with the RSs.

Because the transparent operation is limited to two hops, only one access zone is required in the DL and one pair of access and relay zones are required in the UL. As such, the BS is the only node responsible for sending control information and managing radio resources. All tRSs conform to their superordinate stations, that is, only centralized scheduling is exercised in transparent relay.

4.2.2 Frame Structure in Non-Transparent Relaying

In non-transparent relaying, resources are scheduled in either a centralized or a distributed manner. In centralized scheduling, the BS generates the control information and sends it to its subordinate RSs. The ntRS relays this information to their subordinates at the start of the DL access zone in the subsequent frames through mandatory R-MAPs and R-FCHs. If the relay path is longer than two hops, each ntRS has to generate its own control information, which may differ from those of the serving BS.

Due to the potential scalability issues of centralized scheduling, distributed scheduling provides an alternative whereby scale and fault tolerance can be achieved. An ntRS in distributed scheduling generates its own schedule and transmits its own R-MAP and R-FCH to its neighbors and subordinates.

Since ntRSs are responsible for either some or all of their control messages in both scheduling types, synchronization has a large impact on the performance of the relay cell, that is, frame headers, DL data bursts, and UL data bursts, must be synchronized.

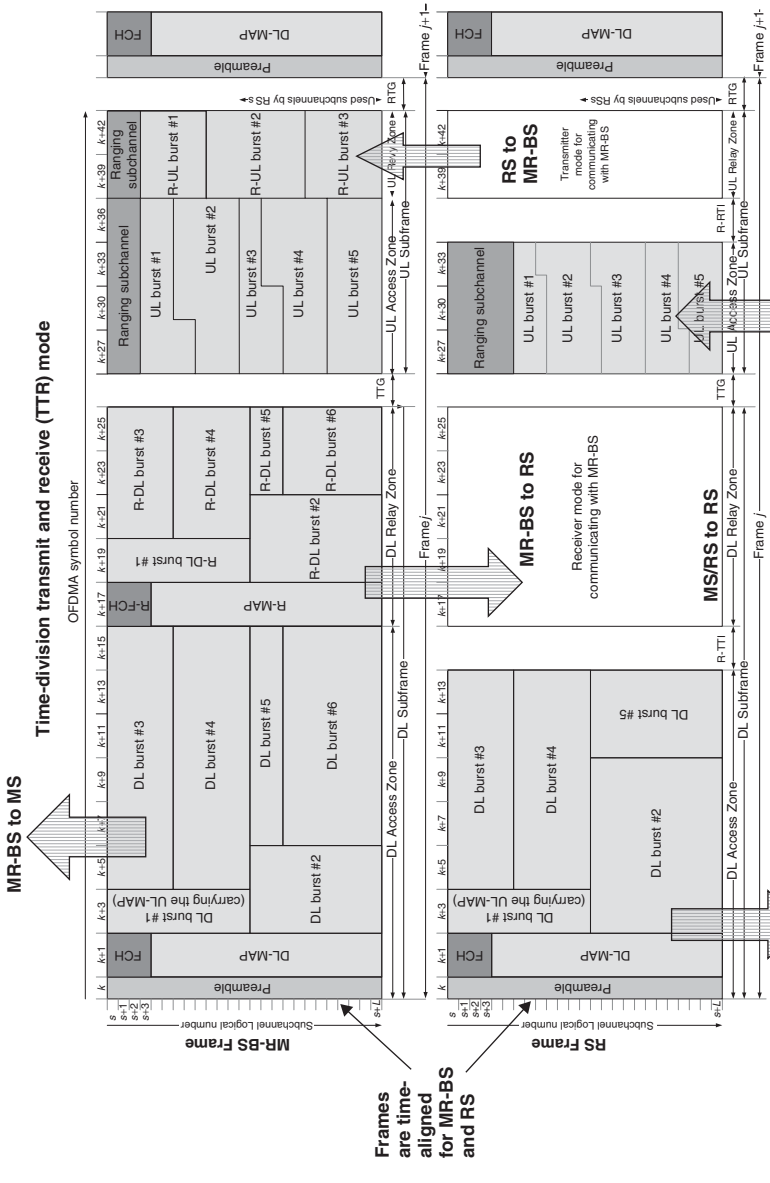


Figure 4.4 Non-transparent frame structure, TTR mode © IEEE 802.16j-2009, Part 16: Air Interface for Broadband Wireless Access Systems Amendment 1: Multiple Relay Specification. Reproduced by permission of © 2009 IEEE.

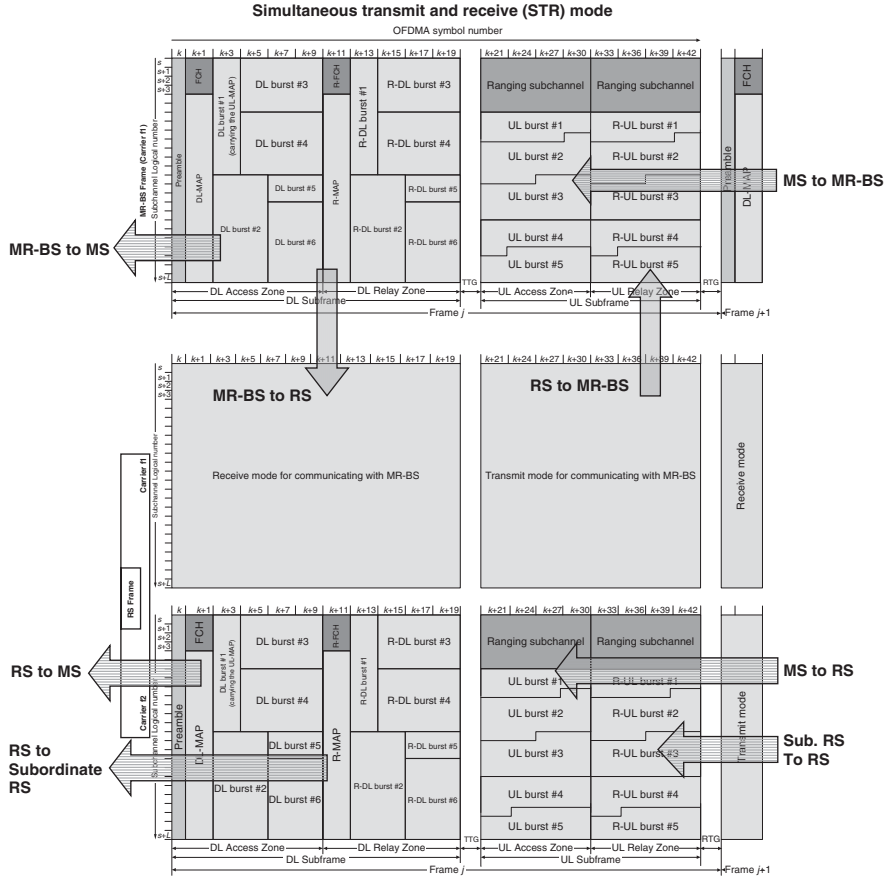


Figure 4.5 IEEE 802.16j-2009, Part 16: Air Interface for Broadband Wireless Access Systems Amendment 1: Multiple Relay Specification. Reproduced by permission of © 2009 IEEE.

There are two possible modes of operation in non-transparent relaying, namely Time-division Transmit and Receive (TTR) and Simultaneous Transmit and Receive (STR). STR allows ntRSs to simultaneously communicate with subordinate and superordinate stations at the same time through using separate radio channels. The frame structure for the TTR and STR in the non-transparent relaying case is shown in Figure 4.4 and Figure 4.5, respectively. A DL subframe in non-transparent relaying must include at least one DL access zone and may include one or more relay zones, while the UL subframe may include one or more UL access zones and one or more relay zone. A relay zone may be utilized for transmitting, receiving, or being idle. A RS, however, shall not be required to support both transceiver modes within the same zone.

The TTR frame structure of the non-transparent mode is similar to that of transparent relay with R-MAP. The difference, however, is that in TTR, both the BS and the RSs can transmit data in the access zone simultaneously due to low interference resulting from the extended coverage, that is, frequency reuse. Meanwhile, the STR frame structure differs from that of the TTR in the usage of dual radio, as shown in Figure 4.5. Accordingly, the STR frame does not require transition gaps between zones to allow the ntRS to switch between transmission and reception.

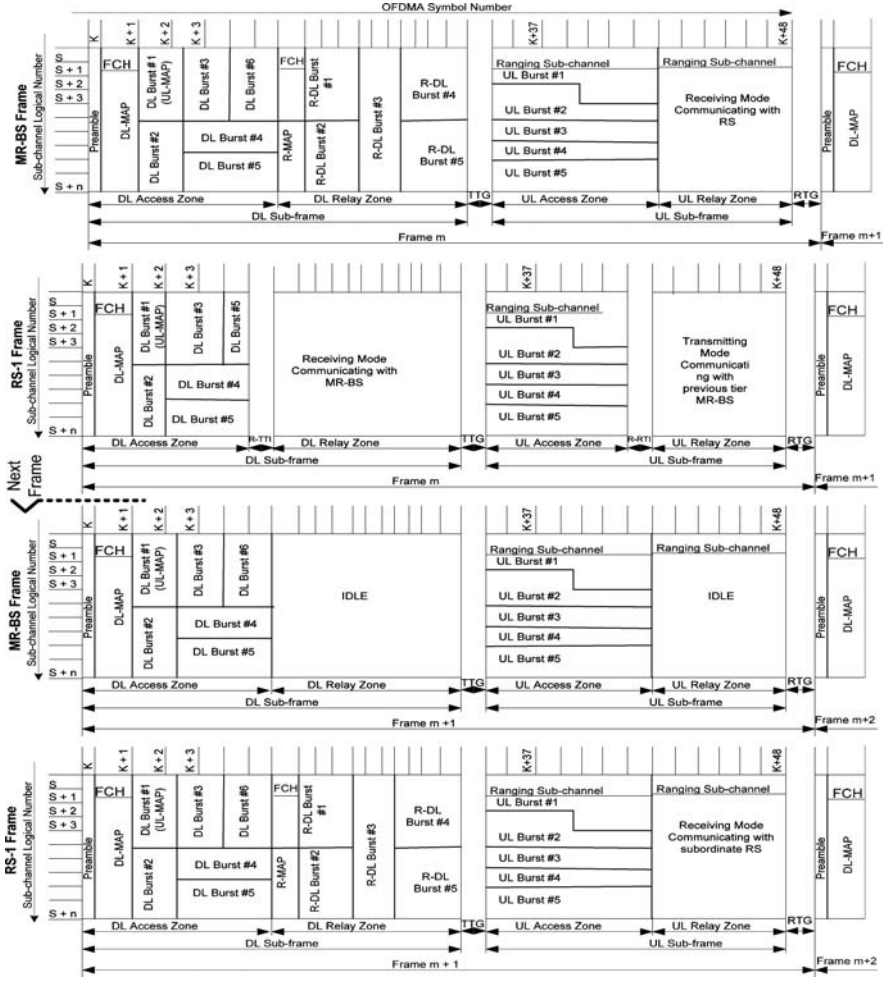


Figure 4.6 Example of non-transparent multi-frame structure, TTR mode.

Non-transparent relaying, especially when TTR is used, allows for more than two hops communications. More than one relay zone is hence to be expected in such topologies. The standard describes two approaches for supporting TTR ntRSs. The first approach groups frames into multiple frames with a repeating pattern for relay zones. An example of such pattern is shown in Figure 4.6. Another example is provided by the standard where two frame sequences are input into a multi-frame, with even numbered hops transmitting in even numbered frames and odd numbered hops in odd numbered frames.

In the second approach, a single frame is constructed to include more than one relay zone. As an example, even numbered hops can transmit in even numbered relay zones, and so on. The single frame may include more than one relay zone in a topology of three hops or more. As an ntRS is only allowed a single transmission, that is, one in the DL and one in the UL, a RS may remain idle for very long durations. This potentially reduces the throughput of the cell. The multiple frame approach, on the other hand, limits the partitioning in the grouped frames, that is, in the single frames of the multi-frame group, an ntRS will be limited to transmit or receive but this may result in an increase in delay. While the standard specifies the single frame structure, it does not describe operation using multiple frames.

4.3 Frame Structure in IEEE 802.16m

To ensure backward compatibility with the IEEE 802.16-2009 frame structure, the amendment proposes a frame structure that achieves two objectives: legacy support for the Reference System and enable the use of new physical and medium access control layers features. Such features, for example, low latency, are not supported in legacy IEEE 802.16.

4.3.1 Basic Frame Structure

The IEEE 802.16m amendment describes a superframe structure of 20ms that supports both TDD and FDD (both full and half). The superframe begins with the superframe header containing control and management information. For backwards compatibility, the superframe is divided into four equally sized 5 ms frames, each consisting of eight subframes for either DL or UL transmission. This frame structure is shown in Figure 4.7 below.

Three types of subframes are defined: type 1 with six symbols, type 2 with seven symbols, and type 3 with five symbols. These types are applied to both FDD and TDD duplexing schemes, including half FDD MS operation. The TDD frame is designed to have two points to switch from DL to UL and vice versa. The half FDD frame structure is similar to that of TDD in employing transmission gaps for switching between DL and UL transmissions. The half FDD

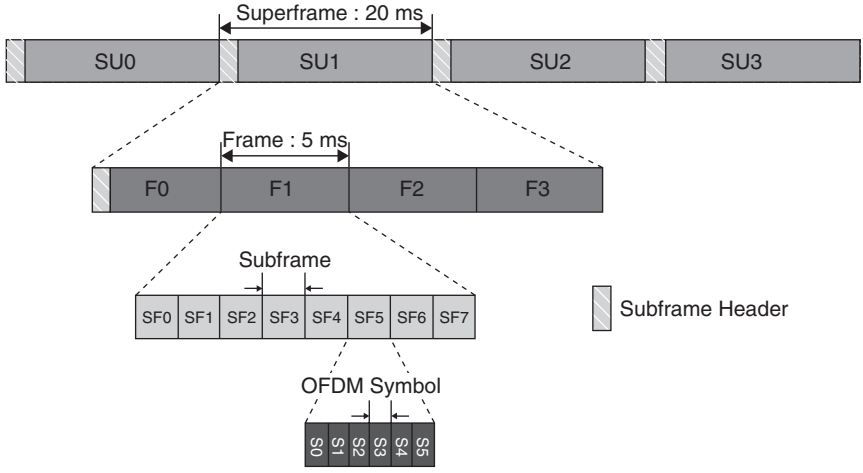


Figure 4.7 IEEE 802.16m basic frame structure. Reproduced by permission of © 2009 IEEE.

frame differs, however, in the DL and UL being transmitted over two separate frequency bands.

4.3.2 Frame Structure Supporting IEEE 802.16-2009 Frames

The above described partitioning of the frame structure into frames and subframes adds flexibility in accommodating a legacy frame within an Advanced frame. Primarily, this is achieved through careful resource allocations. The legacy and Advanced frames, however, are offset by an integer numbers of subframes to accommodate new features as lower latency and smaller control overhead. The supporting frame structure is shown in Figure 4.8.

A different notion of zones, called time zones, is introduced in the IEEE 802.16m amendment, and is applied in both duplexing schemes. A time zone consists of an integer number of adjacent subframes, and is defined to provide support for mixed deployment of Legacy (R1) and Advanced mobile stations. In such a setup, an R1MS is allowed to transmit in a zone called LZone and the AMS is allowed to transmit in both LZone and MZone. The duration of LZone and MZone may vary. An LZone frame starts with a preamble and a MAP, and contains the IEEE 802.16-2009 DL. The UL portion of the frame starts with IEEE 802.16-2009 UL zone to support mixed deployment in the same band and geographical area. The LZone can be removed in pure Advanced deployments.

Figure 4.9 shows the TDD and FDD LZone and MZone under TDM multiplexing. The manner in which time zones is applied for IEEE 802.16m relaying networks is similar to that of the IEEE 802.16j amendment.

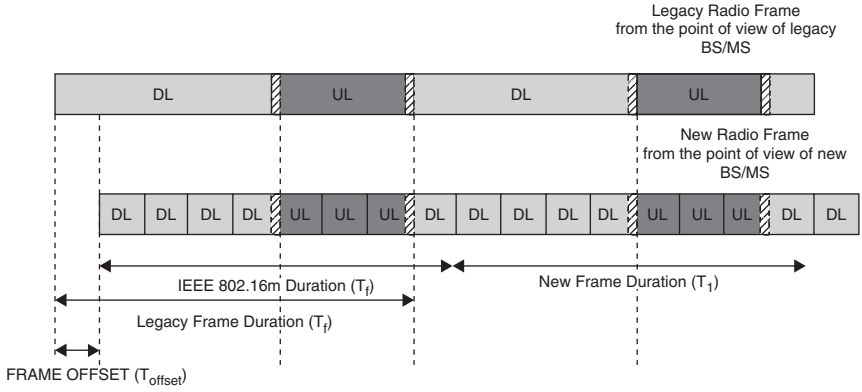


Figure 4.8 IEEE 802.16m frame structure supporting IEEE 802.16-2009 frames. Reproduced by permission of © 2009 IEEE.

4.4 Addressing and Connections Identification

In IEEE 802.16-2009, IEEE 802.16j-2009 and IEEE 802.16m, air interfaces are identified for the different elements (BS SS, RS, and MS) by a unique and universal 48-bit MAC address. This address is not used for identifying the MAC data PDU as in other IEEE 802.x technologies; rather, it is used during the initial ranging and authentication process. The standard also defines logical identifiers to facilitate data and control operations in the different operational modes (PMP, MR).

4.4.1 Logical identifiers in IEEE 802.16-2009

A connection-oriented technology, IEEE 802.16 establishes a logical link between the BS and MS MAC layers that is identified by a 16-bit unidirectional Connection Identifier (CID). The CID is used in the MAC PDU header as a temporary address for the data transmission. Three types of management connections are defined:

1. *Basic*: A mandatory connection established during initial ranging to exchange short, time-urgent MAC management messages;
2. *Primary*: A mandatory connection established to exchange longer, more delay-tolerant MAC management messages; and
3. *Secondary*: An optional connection to transfer delay-tolerant, standards-based messages such as Dynamic Host Configuration Protocol (DHCP), Trivial File Transfer Protocol (TFTP), and Simple Network Management Protocol (SNMP) messages.

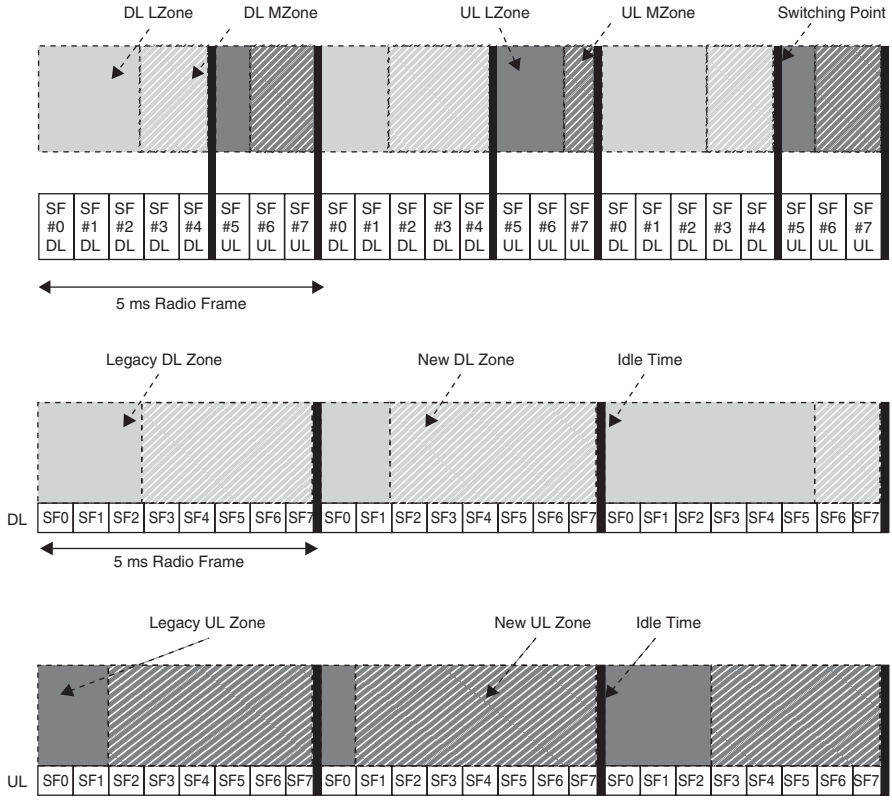


Figure 4.9 Example of Time zones in TDD and FDD modes. Reproduced by permission of © 2009 IEEE.

Another logical identifier is the Service Flow Identifier (SFID) of 32-bit size. A service flow is a unidirectional flow of packets with a specified set of QoS parameters that may be provided through a network management system or created dynamically through defined signaling mechanisms. The BS is responsible for issuing the SFID and mapping it to a unique CID, which in turn is mapped to higher-layer addresses. The CS keeps track of the mapping of each upper layer unit by keeping tracking of several data elements such as the data units destination address, the data units' QoS parameters, source address and SFID and the respective CID.

4.4.2 Logical identifiers in IEEE 802.16j-2009

The IEEE 802.16j amendment defines other logical identifiers that apply to MR mode. The standard requires assigning ntRSs a BS identifier with the same format of the BS identifier defined for IEEE 802.16-2009. Since connections established

may span more than one hop in a relay cell, the CID is unique in the whole relay cell.

The amendment also defines a new connection type, called tunnel, that is established between a MR-BS and access RS. Tunnel connections are setup over a path that may include more than one intermediate RS. Two types of tunnel connections are defined: management, for transferring management PDUs, and transport, for either UL or DL traffic. Management connections are identified by the MT-CID and can be either unidirectional or bidirectional, while transport connections are identified by the T-CID and can only be unidirectional.

4.4.3 Logical identifiers in IEEE 802.16m

An ABS in IEEE 802.16m assigns each AMS and ARS a unicast 12-bit identifier called the Station Identifier (STID). STIDs are assigned during network entry. An ABS may reserve some STID for broadcast, multicast or ranging. The amendment also identifies a 4-bit Flow identifier (FID) to connections. Similar to the CID, an FID identifies management and transport connections. Some FIDs may be pre-assigned. In Advanced relay networks, a tunnel connection is identified by the STID together with the FID.

5

Network Entry, Initialization and Ranging

Both the IEEE 802.16-2009 standard and the IEEE 802.16m amendment define procedures for network entry, initialization and ranging. These procedures configure connectivity parameters, achieve synchronization and enable power control for both MSs and RSs. The IEEE 802.16m extends these procedures through refining the definitions of the different operational states of network entities.

This Chapter is organized is organized as follows. Section 5.1 discusses network entry, initialization and ranging in the point-to-multipoint model of IEEE 802.16-2009. In addition to the OFDMA interface, the section briefly touches on ranging in OFDM given its relevance to understanding the procedure. Section 5.2 elaborates an entry, initial ranging and periodic ranging in the multihop relay amendment, IEEE 802.16j-2009. Finally, Section 5.3 discusses network entry the WiMAX IMT-Advanced amendment, IEEE 80.216m.

5.1 Network Entry in IEEE 802.16-2009

The IEEE 802.16 standard distinguishes ten procedures in network initialization and entry. Of these ten procedures, shown in Figure 5.1, four procedures are implementation dependent and were hence identified as optional. These ten are (optional procedures are marked with an asterisk (*)):

1. Scanning and synchronization;
2. Obtaining downlink parameters;
3. Initial ranging and automatic adjustments;
4. Negotiating basic capabilities;
5. Authorizing SS and performing key exchange*;
6. Registering with BS;

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

7. Establishing IP connectivity*;
8. Establishing time of the day*;
9. Transferring operational parameters*;
10. Establish provisioned connections.

Upon initialization or powering up, an SS scans the band for a downlink channel. Once it recognizes one DL-MAP message and discerns the respective downlink Burst Profile information, synchronization is said to be achieved. An SS

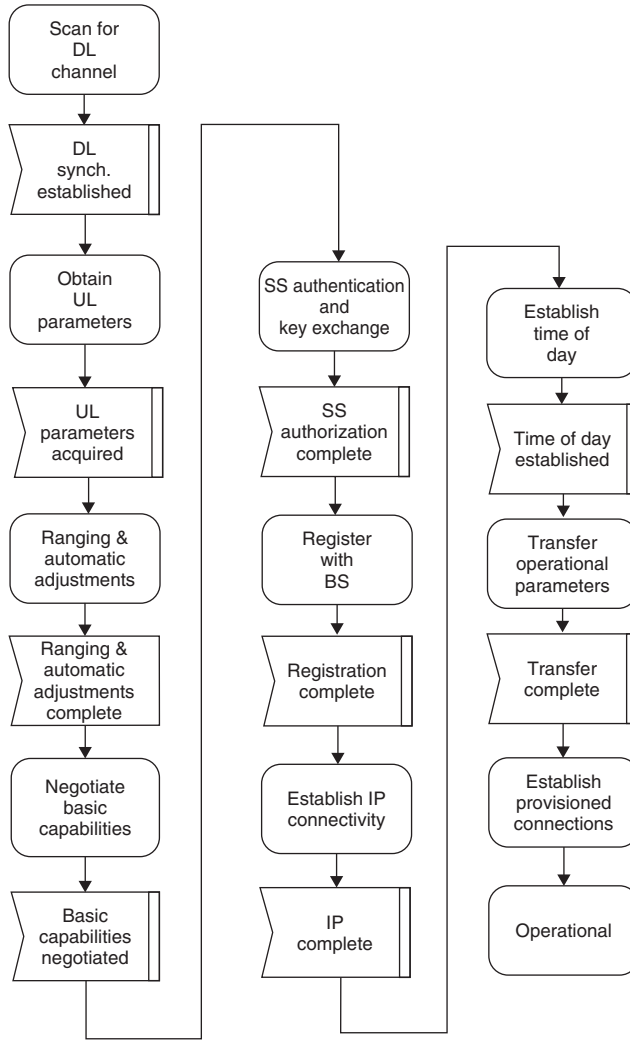


Figure 5.1 Network Entry and Initialization flowchart for IEEE 802.16-2009. Reproduced by permission of © 2009 IEEE.

remains synchronized as long as it continues to successfully receive the DL-MAP and the DCD message for the channel.

Once synchronization is established, the SS waits for an UCD to distinguish a possible uplink channel. A BS periodically transmits the UCD to the MAC broadcast address. If the SS cannot distinguish an uplink channel, it looks for another downlink channel. An SS distinguishes the operating mode (whether FDD or TDD) through differentiating between the center frequency in the DCD and the UCD. The absence of any frequency Type/Length/Value (TLV) in the channel descriptor indicates TDD. Once an SS finds an uplink channel that is most suitable for its purposes, it performs initial ranging.

Initial ranging is the procedure by which the BS recognizes an SS by its MAC address and transmission and reception capabilities. Through initial ranging, as well, the BS adjusts the SS's parameters such as transmission power, time offset, and frequency offset, in order to regulate interference and signal quality within the cell/sector. This procedure is expanded on in next section.

Once initial ranging is completed, the SS proceeds to negotiate basic capabilities with the BS. If authorization is enabled in the network, the SS will perform authorization and key exchange. Registration marks the final major procedure for a non-managed SS in the network entry process. A managed SS would indicate that it is so during initial ranging and, through registration, would obtain its secondary management CID and know which IP version is used. It would then proceed to establish IP connectivity through either Mobile IP or DHCP over the secondary management connection. The optional stages of establishing time of the day and transferring operational parameters (for managed SSs) are then pursued if needed.

The SS could now establish provisioned connections through Dynamic Service Addition (DSA) request and response messages (DSA-REQ and receiving DSA-RSP in response), and is considered operational. An SS's operational status thereafter is maintained through periodic ranging.

5.1.1 Initial Ranging

There are two ranging processes that an SS undergoes: initial and periodic. Initial ranging, made in the initial ranging contention-based interval, is made during two phases of operation: (re)registration or when synchronization is lost, or during transmission on a periodic basis. On the other hand, periodic ranging uses the regular uplink bursts granted by the BS. For a BS, the duration of the ranging slot for initial system access depends on the intended cell radius.

To contend, the SS scans the UL-MAP for initial ranging interval. A BS is required to afford transmission opportunities. For SC-FDMA as well as OFDM – based PHY, the size of each transmission opportunity (TxOP) is specified by the UCD TLV Ranging Request Opportunity Size. Moreover, the SS puts together a Ranging Request (RNG-REQ) in the initial ranging interval. On the other hand, for OFDMA PHY, the SS sends an initial ranging CDMA code on an uplink allocation dedicated to this objective.

There are a total of 256 CDMA ranging codes utilized in OFDMA, each having a length of 144 bits. Each BS is assigned a fraction of these codes, denoted S, and ranging between the values of S to $((S + 0 + N + M + L) \text{ mod}(256))$, where N codes are used for initial ranging, M for periodic ranging, L for bandwidth requests and O for handover ranging. In this manner, the BS can determine the purpose of the received code by identifying the subset to which it belongs.

When an initial ranging interval transmission opportunity occurs, the SS would send the RNG-REQ (or CDMA code in OFDMA). The SS then sends the messages as if it was collocated by the BS.

The SS sends its RNG-REQ at a power level below the maximum allowed ranging transmission power ($P_{TX\_IR\_MAX}$). If the SS does not receive any response, it adjusts its power level until success is indicated by an RNG-RSP with the SS's MAC address. A BS that is unable to decode an RNG-REQ would send an RNG-RSP with only the request's transmission parameters and indicating frame opportunity. For OFDMA, the SS sends the CDMA code at a power level below $P_{TX\_IR\_MAX}$ and would increase the power level if no response is received. An unsuccessful attempt would be indicated by the BS via an RNG-RSP with code parameters and a Continue status. The SS would then implement corrections and randomly select the next ranging slot. An UL-MAP with a CDMA allocation IE containing the SS's code parameters is considered an RNG-RSP indicating success. When received, the SS sends an RNG-REQ in the indicated bandwidth slot.

An RNG-RSP would be identified by the SS's initial ranging CID. It would also contain the Basic and Primary Management CIDs. If needed, the RNG-RSP would also include adjustments on the SS's RF power level and offset frequency, in addition to corrections in the timing offset. An RNG-RSP with a Success status indicates the end of the initial ranging process. An RNG-RSP with Continue status will make the SS wait for an individual initial ranging interval assigned to its Basic CID for its next RNG-REQ. Once an SS is "ranged" it joins the BS's normal data traffic.

It is possible during network entry that a BS redirects a ranging SS to another channel with an offset frequency adjustment. If the adjustment is less than half the channel's bandwidth, it would be considered as a fine tuning. However, if the adjustment is greater than half the channel's bandwidth, the SS would understand that this is effectively a channel reassignment and would have to restart the ranging process on the new channel.

Any adjustment made by the BS must be within the standard's defined operating ranges. An SS response to an RNG-RSP, including any required adjustments, is mandatory. An SS will not make any transmission until adjustments indicated in an RNG-RSP are made.

5.1.2 Periodic Ranging

To maintain connection quality for an SS, the BS and SS are engaged in a periodic ranging. Distinct ranging processes are used for managing the downlink and the

uplink. At the same time, certain PHY modes support ranging mechanisms unique to their properties. In what follows, periodic ranging in OFDM and OFDMA PHY is explained.

5.1.3 Periodic Ranging in OFDM

The signal quality at an SS determines the selection of the burst profile at the BS. To reduce uplink traffic volume, an SS monitors the Carrier-to-Interference-Noise-Ratio (CINR) it perceives and compares the average value against the allowed range. If the SS finds the CINR out of preset bounds, the SS requests a change of burst profile by either using the allocated data to send a downlink Burst Profile Change Request (DBPC-REQ) or starting an initial ranging. The latter option is used only when the SS is interested in a more robust profile. A BS receiving a DBPC-REQ shall respond with a DBPC-RSP indicating whether a change in the SS's burst profile is possible.

For uplink ranging, the BS maintains a timer (T27) for each SS that resets whenever a unicast grant is made. Upon expiry, a BS grants bandwidth to the SS for an uplink transmission in the form of a data grant or an invited ranging opportunity. The SS maintains another timer (T4) indicating how long has it been since the SS was given an opportunity to transmit. Once this timer expires, the SS restarts all its MAC operations. In turn, the BS monitors an SS's use of its unicast grants and terminates the link if it has not been utilized for certain duration. For each utilized uplink grant, the BS adjusts the SS's power level through the use of an RNG-RSP to which the SS must adhere. If the SS does not make the required adjustments, the BS would terminate the connection through the use of an aborting RNG-RSP.

An SS interested in maintaining its connection would always utilize its uplink data grant with either data, an RNG-REQ, a padding PDU or stuff bytes.

5.1.4 Periodic Ranging in OFDMA

Periodic ranging in OFDMA utilizes a regular uplink burst. The ranging channel is composed of one or more groups of six adjacent subchannels. Groups are defined starting from the first subchannel, and channels are considered adjacent if they have successive logical numbers. The indices are specified in the UL-MAP and users are allowed to simultaneously contend and collide.

For OFDMA, the standard specifies both the ranging subchannels and special pseudonoise ranging code. As explained above, different subsets of the code are used for different objectives. For each objective, an SS would select (with equal probability) one of the codes from the respective subset, modulate it onto the ranging subchannel choosing (with equal probability) a slot from the available ranging subslots. When needed, backoffs with random duration are used to mitigate contention. As the BS cannot identify SSs through code alone, a BS

broadcasts ranging response with code, ranging slot that had the code (OFDMA symbol number, subchannel), and required adjustments (time, power, frequency).

5.2 Network Entry in IEEE 802.16j-2009

An MS under an MR network would undergo the same procedures for network entry, initialization, and periodic ranging as it would in a PMP network. To achieve this compatibility, there are operations to be noted by the MR-BS and the RS in order to maintain a non-MR activity for the MS. As such, operations made on part of an MS during these procedures will not be discussed in this section. The scope of this section hence will comprise two sets of descriptions, a description of the actions taken by an MR-BS to carry out MS procedures, and a description of the procedures required for RS operation. An RS naturally requires additional steps to become fully operational. Figure 5.2 schematizes the network entry and initialization procedures for an RS. The standard distinguishes eleven major steps, with four additional sub-procedures required for RS. In the following list, the optional procedures are marked with an asterisk (*). A distinction is also made between the procedures required for SS only or RS only.

a. Scanning and synchronization;
   a1. Performing first stage access station selection (RS only);
b. Obtaining transmission parameters from UCD message;
c. Ranging;
d. Negotiating basic capabilities;
e. Authorizing SS/RS and performing key exchange*;
f. Performing registration;
   f1. Obtaining neighbor station measurement report (RS only)*;
   f2. Performing the second station access selection (RS only)*;
   f3. Path creating and tunnel establishment (RS only)*;
g. Establishing IP connectivity*;
h. Establishing time of day*;
i. Transfer operational parameters*;
j. Setting up connections (SS only);
k. Configuring operation parameters (RS only).

In addition, in an accelerated network entry operation for RSs called RS network entry optimization, an MR-BS may instruct the RS to omit procedures d–f (1–3), and k.

RSs follow the same scanning and synchronization procedures as SSs. An RS may optionally store preamble indexes and corresponding signal strengths, and report them if requested by the MR-BS during procedure f1.

To assist the first stage access selection, an MR-BS and operating RSs may transmit a TLV with an end to end metric in the DCD. This metric would be

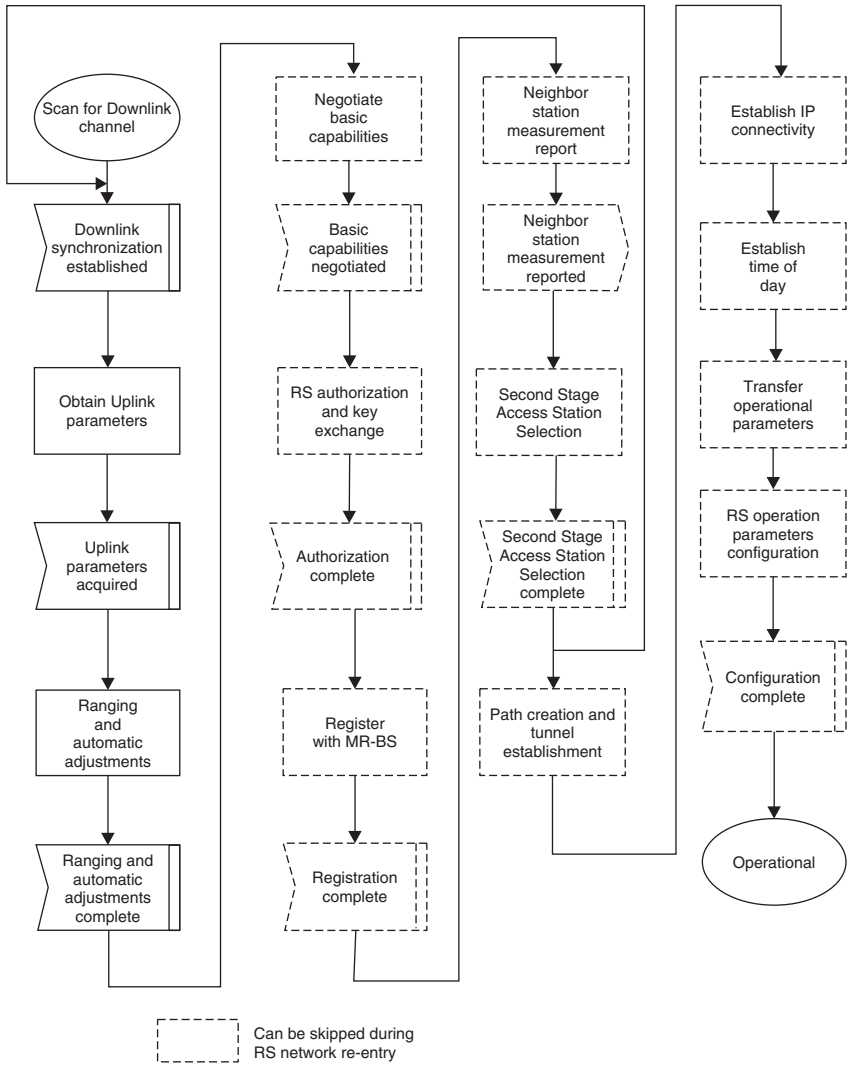


Figure 5.2 Network Entry and Initialization flowchart for IEEE 802.16-2009. An adaptation from Figure 65a, page 102 in 16j-2009 "RS Initialization". Reproduced by permission of © 2009 IEEE.

considered in an entering RS's decision to select an access station. Once an access station is selected, the RS carries on with the network entry procedure.

MR-BSs and RSs involved in registration shall perform as described in PMP. However, if the MR-BS decides that neighbor measurement report is required, the entering RS shall provide the report with the requested details. Once the RS has sent a REG-REQ to the MR-BS, it shall wait for a REG-RSP.

A neighbor station measurement report can include the signal strengths and the respective preamble indexes of neighboring ntRSs with unique BSIDs or signal strength and R-amble indexes of neighboring tRSs or ntRSs with shared BSIDs. When measurements are not required, the RS skips both reporting and the second stage access selection and proceeds to the next procedure indicated in the MR-BS's RNG-RSP message. Note that mobile RSs may be instructed as to which preamble indexes should they collect the measurements for. If an RS is requested to report measurements, the MR-BS shall not change the frame configuration for the RS's superordinates before the RS becomes fully operational.

The purpose of the optional second stage access station selection is to direct the entering RS to another access station for network and resource management objectives like interference management or load balancing between relay paths. This procedure depends on the neighbor measurement report. If the entering RS is to associate with the access station selected in the first stage, it shall proceed with remaining network entry procedure. If the current access station is changed to another that is under the same MR-cell, the MR-BS shall indicate the new station's preamble index to the RS, and both the MR-BS and the RS will be engaged in a network reentry procedure (for which network entry optimization may be applied). If the network reentry fails, the originally selected access station shall be used as a first candidate for reentry.

The optional path creation and tunnel establishment procedure can be used to create a path, establish tunnels or bind tunnels to an already active path, and is performed after an RS successfully completes the process for access RS selection. Single tunnels can only be used for either management or transport.

5.2.1 Initial Ranging

As the SS carries on the initial ranging procedures described above, the MR-BS and the RS need to manage the relevant signaling in a seamless manner with respect to the SS. Subtle differences in behavior largely depend on network configuration, that is, whether SS is performing initial ranging through a tRS or an ntRS, whether the implemented scheduling is centralized or distributed, and whether a group of RSs share a BSID or each RS has its own unique BSID. As will be noted below, an RS initial ranging procedure closely resembles that of an SS except for certain modifications.

In a network where a tRS is connected directly to an MR-BS, the RS monitors the ranging channel on the access link for initial ranging codes. The codes are then relayed to the serving MR-BS with proper adjustments (for time, power, etc) in an MR_RNG-REP. The MR-BS, in turn and after waiting for MR_RNG-REP from other stations, decides on the best path for the SS. If adjustments are required, the MR-BS sends an RNG-RSP to the SS. Otherwise; the MR-BS makes an allocation in the access uplink for the SS so that it would send its own RNG-REQ.

If the tRS is attached to the MR-BS through a centralized ntRS, the tRS monitors the ranging channel in the UL-MAP set by its superordinate station.

The ntRS then manages the tRS's MR_RNG-REP, schedules downlink allocation to send the RNG-RSP to the SS and, upon requiring no further adjustment, request uplink bandwidth for the SS to send its RNG-REQ.

In instances where the SS is performing initial ranging with a group ntRSs sharing a BSID, the MR-BS needs to decide whether to specify the access RS as a receiving, or utilize the multicast management CID for the shared BSID group.

When an SS is performing initial ranging through a tRS attached to a scheduling RS, or through ntRSs sharing a BSID, the scheduling RS shall perform adjustments directly with the SS without getting back to the serving MR-BS. The scheduling RS will also independently manage bandwidth allocations for relaying MAC messages to and from the SS.

For ntRSs with centralized scheduling and unique BSIDs, the ntRS monitors the ranging channel. Upon receiving a ranging code, the ntRS shall determine whether adjustments are necessary and, if they are, the ntRS will seek allocation for the RNG-RSP from the MR-BS. If no adjustments are required, the ntRS shall request an uplink allocation for the SS to send its RNG-REQ. The RNG-REQ is ultimately handled by the MR-BS.

When an SS is dealing directly with a scheduling RS, the RS monitors the ranging channel specified in its own UL-MAP. When the RS detects a ranging code on its access link, it shall perform adjustments directly with the SS.

Finally, when an RS is performing initial ranging, it shall follow the procedures described above for the SS except that it will use an RS initial ranging code instead of a regular initial ranging code. After receiving an RS initial ranging code, the MR-BS or the ntRS may send an RNG-RSP indicating preamble indexes of candidate neighbor stations. In all, operating tRSs ignore an RS initial ranging codes.

Note that, similar to IEEE 802.16-2009, the CDMA ranging codes utilized in OFDMA are 256 codes, each consisting of with 144 bits. Each BS is assigned S codes from the 256, that is, the range between S to (($S + O + N + M + L + P + Q$) mod(256)) where O, N, M *and* L are used for normal ranging and P and Q for RS initial ranging and RS unique CDMA ranging, respectively.

5.2.2 Periodic Ranging

Periodic ranging for both SSs and RSs proceeds as in periodic ranging for PMP operation.

For an SS, the ranging completes when the access station (whether an MR-BS or an ntRS) sends the RNG-RSP. For an RS, an MR-BS may assign a dedicated RS CDMA periodic ranging code. Again, the ranging process completes once the access station to which the RS is attached sends an RNG-RSP.

Superordinate stations to an SS/RS may initiate periodic ranging based on measurements. This involves sending an unsolicited RNG-RSP. If the superordinate is a tRS, the MR-BS and tRS shall proceed normally. If the superordinate is tRS or ntRS in an RS group, an MR_RNG-REP is sent to the MR-BS to request that an RNG-RSP be sent (with the necessary adjustments) to the SS.

If the superordinate is a centralized ntRS, it will seek the MR-BS for downlink allocation to send the RNG-RSP. Finally, if the superordinate is a scheduling RS, it would send the RNG-RSP directly without going back to the MR-BS.

5.3 Network Entry in IEEE 802.16m

Under IEEE 802.16m, an AMS transitions between four states that are shown in Figure 5.3. These states are:

- Initialization State;
- Access State;
- Connected State; and
- Idle State.

The network entry and initial ranging procedures are performed while an AMS is in Initialization and Access States. Periodic and other (e.g., for handover) ranging procedures are performed while an AMS is in a connected state.

The procedures can be related to the states as follows. When an AMS is in the initialization state it performs the following tasks:

a) Scanning and synchronization;
b) Super-Frame Header Acquisition;
c) Cell selection decision;

Once b) and c) are performed, the AMS transitions to the Access State and performs the following tasks.

d) Ranging and uplink synchronization;
e) Pre-authentication capability negotiation (if applicable);
f) MS authentication, authorization and key exchange;
g) Capability exchange and registration with serving ABS;
h) Initial service flow establishment.

An AMS transitions to the Connected State once h) is performed.

For multicarrier operation, an AMS only attempts network entry and initial ranging with a fully configured carrier. Detecting an A-PREAMBLE, the AMS would decode the SFH and other system parameters and configuration information by which the ABS indicates its support for the multicarrier feature. Once a candidate primary carrier is selected by the AMS, network entry proceeds normally. If successful, the same carrier becomes the primary carrier for the AMS through which it may negotiate with the ABS parameters for secondary carriers. Uplink ranging may be skipped for the secondary carrier if the AMS can utilize configuration information for time, frequency and power configuration from the primary carrier, and may use adjustments in the primary carrier as an initial configuration. If applicable, an ABS may enhance the AMS's ranging in secondary

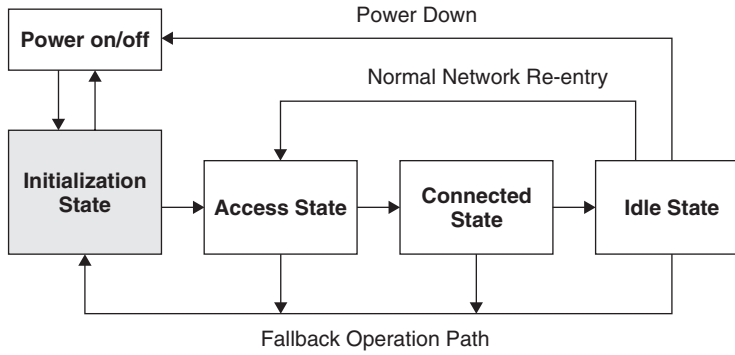


Figure 5.3 Transitions of an Advanced Mobile Station (AMS).

carriers through assigned dedicated ranging codes. These dedicated codes would be communicated to the AMS using the primary carrier.

A Femtocell BS handles AMS entry and ranging procedures similar to a regular ABS. However, an AMS shall not attempt network entry (or handover, reentry from idle, or location update) to a CSG-Closed Femtocell BS except in case of emergency. Similarly with CSG-Open Femtocell BS, an AMS may only attempt such operations if it is critical to the AMS's operation, that is, the AMS connection would be otherwise terminated. An OSG Femtocell BS, on the other hand, is fully accessible to all AMSs within the Femtocell's coverage. Differentiation could be made in ranging contention for CSG-Open Femtocells, whereby CSG-members would always be granted priority over CSG-nonmembers.

An ARS transitions between three states.

- Initialization State;
- Access State;
- Operational State.

Network entry and initial ranging are performed between the Initialization and Access States, while periodic and other ranging types are performed in the Operational States.

In the Initialization State, the ARS performs the following procedures

a) Scanning and downlink synchronization (A-Preamble Detection);
b) Broadcast channel acquisition;
c) Access station selection decision;

The completion of procedure c) enters the ARS into the Access State, in which the following procedures are performed.

d) Ranging and uplink synchronization;
e) Pre-authentication capability negotiation (if applicable);

f) ARS authentication, authorization and key exchange (if applicable);
g) Capability exchange and registration with servicing ABS;
h) Neighbor station measurement report and access station selection (if required by the ABS);
i) Configuring ARS operation parameters.

The completion of procedure i) enters the ARS into the Operational State. It is hence apparent that an ARS follows the same procedures as an AMS. An ARS, however, may additionally perform interference measurement of neighboring stations (if required by the BS), path creation, and tunnel connection establishment with the ABS.

AMS network entry may be distributed between the ARS and the ABS. This includes procedures such as capability negotiation, connection establishment, authentication and registration. Initial link adaptation is handled by the ARS.

6

Quality of Service and Bandwidth Reservation

Four distinct elements are required to provide an effective QoS support in any network. These are:

1. QoS performance measures;
2. Classification;
3. Signaling bandwidth requests and grants; and
4. Bandwidth allocation and traffic handling.

Both IEEE 802.16-2009 and the IEEE 802.16m amendment provide mechanisms that establish these elements at the physical layer and at medium access control layer. In this chapter, we shall have a detailed look at these mechanisms starting with the IEEE 802.16-2009.

This chapter is organized into three sections with the first describing QoS in IEEE 802.16-2009, the second in the IEEE 802.16j-2009 amendment and the third in the IEEE 802.16m amendment. The organization of the individual sections is almost the same, going through a discussion of the bear classification and the signalling require for making bandwidth requests and relaying bandwidth grants. Details for service flow creation, management and deletion then follows. Descriptions of how bandwidth allocations are made and how traffic transmission errors are handled conclude the chapter. Section 6.1, however, differs in defining the QoS performance measures, which apply for both the IEEE 802.16-2009 and its amendments. Meanwhile, Section 6.2 elaborates on differences in bandwidth allocation and handling when IEEE 802.16 relay stations are employed.

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

6.1 QoS in IEEE 802.16-2009

6.1.1 QoS Performance Measures

The performance level of a connection is normally expressed in terms of its throughput, delay, jitter, priority and packet loss. However, the standard specifies another set of parameters to be used in setting up and maintaining a connection. In the following, the mapping between these two sets of parameters is detailed.

6.1.1.1 Throughput

- *Maximum sustained traffic rate*: This is the peak information rate expressed in bits per second to which, the users' traffic shall on average be policed to conform to. However, this parameter only specifies a traffic bound, not a guarantee that the rate is actually available. It should also be mentioned that the standard does not specify any traffic policing mechanism.
- *Maximum traffic burst*: Is the maximum burst size accommodated for a particular service measured in bits. It is also the maximum continuous burst accommodated by the system for a service if this service is not currently using any of its allocated resources. The maximum sustained traffic rate and the maximum traffic burst are jointly identified by a six-bit code word. The standard documents define twenty three different levels for the maximum traffic burst and maximum sustained traffic rate are defined.
- *Minimum reserved traffic rate*: This last parameter represents the minimum rate reserved for a service flow measured in bits per second. A connection mapped to a certain superframe may request a data rate up to its minimum reserved traffic rate which should be guaranteed by the BS. However, if the requested rate is less than this minimum value, the BS is still required to guarantee the requested rate.

6.1.1.2 Delay

- *Maximum latency*: Specifies the maximum interval, measured in time units, between the reception of a packet (at either the BS or the MS) and the forwarding of the Service Data Unit (SDU) to the air interface. If specified, the value of this parameter will be guaranteed by the network.

6.1.1.3 Jitter

- *Tolerated Jitter*: Specifies the maximum delay variation for a connection in seconds. If specified, this parameter should be guaranteed by the BS. Both maximum latency and tolerated jitter are identified by a six-bit code word.

6.1.1.4 Priority

- *Traffic priority*: Specifies the priority of the associated service flow.

Of these parameters, the minimum reserved traffic rate, the maximum latency, and the tolerated jitter are hard parameters while the maximum sustained traffic rate as well as the traffic priority are soft parameters whose satisfaction depends on the state of the network. Hard parameters are ones that, if accepted by the network, will be guaranteed at the specified values, while soft parameters are one that are either statistically satisfied, or to be satisfied at the network's best effort.

6.1.2 Classification

Classification is the process of mapping a MAC SDU to a particular transport connection for transmission between MAC peers. A transport connection is in turn associated with a service flow that defines the appropriate QoS constraints of the transport connection, and determines the level of treatment the SDU receives from the network.

Each packet entering the network is classified and associated with a connection and an SF based on certain classification rules. Each of these rules comprises a group of criteria upon which the match (packet to connection and SF) is made. The criteria may include destination and source addresses, rule priority (t resolve conflicting rules), and a reference to a unique CID. Several classification rules may refer to the same service flow. However, a packet not matching any CID will be discarded. Figure 6.1 and Figure 6.2 illustrate the mapping of packets to connections and service flows in both, the downlink and the uplink.

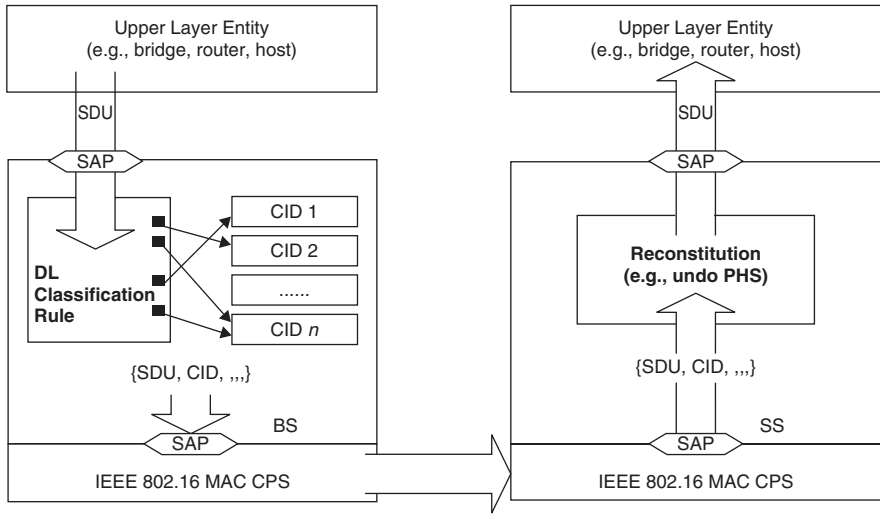


Figure 6.1 Classification and CID mapping (BS-to-MS). Reproduced by permission of © 2009 IEEE.

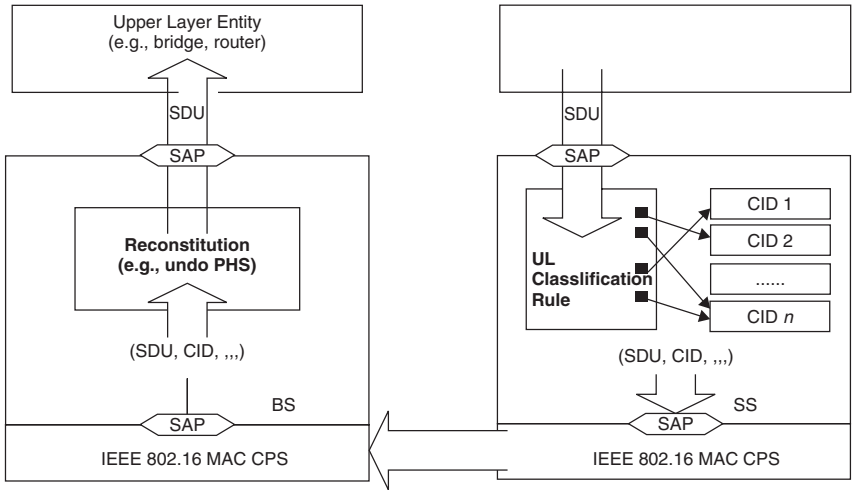


Figure 6.2 Classification and CID mapping (MS-to-BS). Reproduced by permission of © 2009 IEEE.

Service flows can be classified into three types depending on its relationship to the connection. These are:

1. A *provisioned service flow* is associated with the ProvisionedQoSParamSet parameter. This flow is only provisioned and has no connections associated with it. A provisioned service flow should first be admitted to be associated with a connection.
2. An *admitted service flow* is associated with the AdmittedQoSParamSet parameter, which is used by the BS to allocate resources to the service flow based on the contracted QoS parameters. These parameters may include the maximum sustained traffic rate, the minimum reserved traffic rate, traffic priority, tolerated jitter, and the maximum latency parameters. However, not all of these parameters are defined for each flow. For example, tolerated jitter and maximum latency are not defined for non-real time traffic. Moreover, since an admitted service flow is allocated resources by the network; it is associated with an admitted connection.
3. An *active service flow* is an admitted service flow with an active connection which has packets to be transmitted, and is associated with the ActiveQoSParamSet parameter.

Figure 6.3 shows the above noted relationships between the different service flow types, and how a connection should progression from being provisioned, to admitted, to become an active service flow.

The process of activating an admitted services flow is called the two-phase activation model. In the first phase, resources are allocated to the admitted flow.

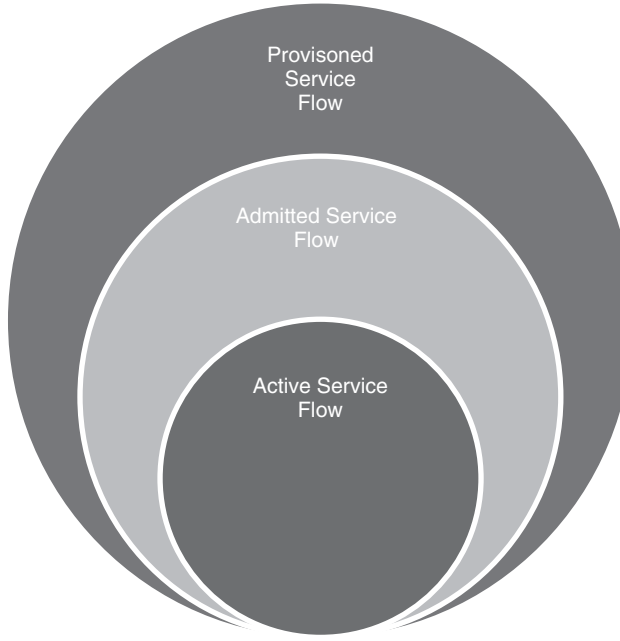


Figure 6.3 The relationship among the different service flow types.

However, these resources are not utilized until ready-to-transmit packets are available, which is the second phase. Consequently, this model conserves the network resources and only uses it when an end-to-end connection is established.

Connections and service flows are used for unicast, multicast and broadcast traffic. A connection and a service flow are required to be admitted before being used for traffic transport. After admitting a connection or a service flow, it can be either altered or deleted. Alteration can be made by requiring a change in bandwidth. An authorization module, a logical module in the BS, processes such requests. The module operates in two modes: static and dynamic. In the static mode, the module oversees the provisioned service flows, maintaining their status and admitting provisioned service flows only if the admitted QoS parameters are a subset of the provisioned QoS parameters. In a similar fashion, from a provision service flow to become an active service, the active QoS parameters must be a subset of the provisioned QoS parameters sets. On the other hand, when the module operates in the dynamic mode, it connects to a policy server that is referred to when processing any admission or activation requests. Based on this referral, which validates whether the AdmittedQoSParamSet or ActiveQoSParamSet is a subset of the set provided by the policy server, requests can be either accepted or rejected.

Service flows can be aggregated based on their classes. A service class is identified by a unique set of QoS requirements. Connections within a certain

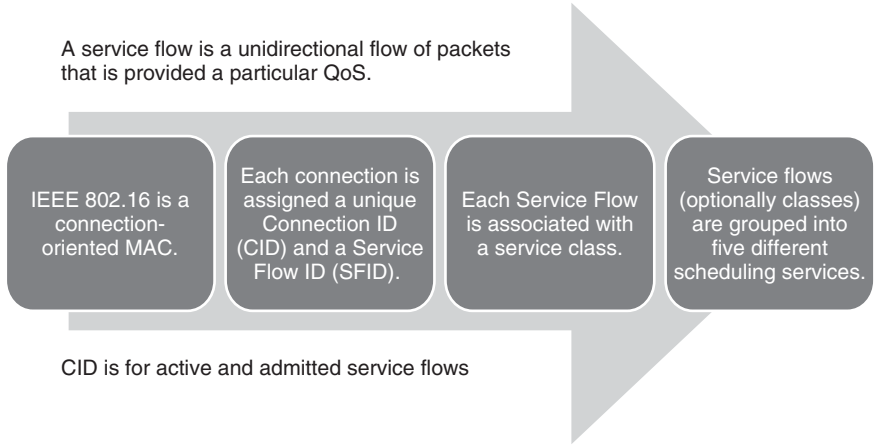


Figure 6.4 Mapping traffic into transport connections, service flows and scheduling services.

class can have individually different QoS requirements. Nonetheless, the standard neither defines specific classes nor limits its number that can be implemented in a network. Hence, using the classes as well as the flexibility of configuring it at the BS is left to network operators.

Grouping services flows (or service classes) into different scheduling services uniquely determines the mechanism by which traffic is allocated to each flow. Figure 6.4 shows the mapping of a traffic flow packets on to a transport connection, a service flow, and a scheduling service.

The standard defines five scheduling services. The services, summarized in Table 6.1, are as follows:

1. *Unsolicited Grant Service (UGS)*: Supports real-time service flows with fixed-size data packets such as T1/E1 and VoIP without silence suppression. The service offers fixed-size grants on a periodic basis, eliminating overhead and latency and assuring that grants are available to meet the flow's real-time requirements. The BS provides bandwidth grants based on the Minimum Reserved Traffic Rate of the service flow, which essentially is the Maximum Sustained Traffic. An MS with a UGS flow is not allowed to use contention slots to request transmission opportunities for this type of services. The standard, however, allows an MS to set a Slip Indicator bit (SI) flag once a UGS flow has exceeded its transmission queue depth. Once the BS receives the flag, it provides for additional grants in the coming frames to compensate. An MS may also use two fields called the Frame Latency (FL) and Frame Latency Indicator (FLI) in the grant management subheader to alert the BS of inordinate latency. Once such fields are noted by the BS, it may schedule earlier grants for the relevant service flow.

2. *Real-time Polling Service (rtPS)*: Supports real-time service flows with variable-size data packets, such as Moving Pictures Experts Group (MPEG) videos. The BS periodically allocates unicast request opportunities for the SSs to request transmission opportunities for this scheduling service. The request opportunities allow an MS to specify the QoS parameters of a desired grant to meet its flow's requirements. The BS may accept or deny an MS's request, for example, based on network capacity. MSs are only allowed to use the unicast request opportunities or piggyback request to a data PDU, and are hence prohibited from using any contention request opportunities for this type of service.
3. *Extended rtPS*: Supports transport variable-size data packets, such as VoIP with silence suppression. Extended rtPS is similar the UGS in having unicast grants in an unsolicited manner, while it resembles rtPS by having periodic unicast request opportunities. Another similarity with the rtPS is that an MS with an enhanced rtPS service flow may piggyback its bandwidth request to a data PDU. Unlike UGS, however, enhanced rtPS results in allocations with variable sizes. An MS with an enhanced rtPS may also contend for a request opportunity, or send a Channel Quality Indicator Channel (CQICH) code word to inform the BS of having data to send.
4. *Non-real-time polling service (nrtPS)*: Supports delay-tolerant services that require allocations on regular basis, such as FTP. Similar to rtPS and enhanced rtPS, the nrtPS service offers unicast bandwidth request opportunities on a regular basis, which assures that an MS may receive request opportunities even during network congestion. However, the BS typically polls nrtPS connections on an interval longer (by one second or less) than that of either rtPS or enhanced rtPS. Unlike the rtPS service, an MS with nrtPS service is allowed to use contention request opportunities.
5. *Best effort (BE) service*: Supports BE traffic and provides little or no QoS guarantees. The BE service has the lowest priority in a network. An MS with BE service contends in the bandwidth contention region to send its bandwidth request to the BS. The BS fulfills the request only if resources are available and are not required by any other scheduling services.

6.1.3 Signaling Bandwidth Requests and Grants

The BS is the sole entity responsible for bandwidth allocations. With data to send, the BS schedules the physical layer resources required to meet the data QoS requirements on a per connection basis. Allocations made to a specific MS are indicated in the DL-MAP over dedicated management connections. Moreover, the BS may increase or decrease a downlink connection's allocations at its own discretion. This applies for all the scheduling services mentioned above except for UGS connections. For uplink connections, however, adjustments are made after processing an MS's request.

Data traffic is transmitted over transport connections, while management messages are transmitted over management connections. These two connection types

Table 6.1 Scheduling services provides a summary of the QoS scheduling services

Service	Definition	Example applications	QoS parameters
UGS	Real-time data flows with fixed-size data packets requiring periodic allocations	VoIP w/o silence suppression	Maximum Sustained Traffic Rate Maximum Reserved Traffic Rate Maximum Latency Tolerated Jitter Request/Transmission Policy
ertPS	Real-time data flows with variable sized data packets requiring periodic allocations	VoIP w/ silence suppression	Maximum Sustained Traffic Rate Minimum Reserved Traffic Rate Maximum Latency Request/Transmission Policy
rtPS	Real-time data flows with variable-size data packets requiring periodic allocations	MPEG Video	Minimum Reserved Traffic Rate Maximum Sustained Traffic Rate Maximum Latency Traffic Priority Request/Transmission Policy
nrtPS	Delay and jitter-tolerant data flows with variable-sized data packets for a which a minimum data rate is required	FTP	Minimum Reserved Traffic Rate Maximum Sustained Traffic Rate Traffic Priority Request/Transmission Policy
BE	Data flows with little or no QoS requirements	HTTP	Maximum Sustained Traffic Rate Traffic Priority Request/Transmission Policy

are associated to service flows as discussed in the previous section. Prior to discussing the bandwidth request and grant mechanism, it is instrumental to discuss the process of creating, managing and deleting a service flow.

Service flows may be provisioned through a network management system, that is, static or configured service provisioning, or created dynamically through defined signaling mechanisms. An MS is neither allowed to alter the bandwidth

requirements of a static provisioned service flow nor to create new static service flows. Policies and entities that dictate such policies are outside the standard's scope, but are considered to be overseen by upper layer management functionalities.

In dynamic service provisioning, the standard defines three processes for the creation, changing, and deletion of a service flow. Each process is carried out with a three-way handshake message delivery. Dynamic Service flow Addition (DSA) is used for creating a service flow, Dynamic Service Change (DSC) is used for changing a service flow and Dynamic Service flow Deletion (DSD) is used to delete a service flow. These are discussed in more details next.

6.1.3.1 Service Flow Creation, Management and Deletion

Either the BS or the MS may initiate a service flow creation process based on the flow's direction, that is, whether downlink or uplink. An MS initiates the creation of a service flow by sending a DSA-REQ message to the BS. The message contains the QoS parameters set and is marked by the type of the service flow (admit only or admit and activate). This exchange is shown in Figure 6.5(a). The BS sends a DSX-RVD message if the integrity of the message is intact. The BS then checks whether the MS's request is admissible and whether the request's QoS parameter set can be supported. If the flow is rejected due to being unauthorized, the BS indicates that in the DSA-RSP message and sends it to the MS. If the flow is rejected due to insufficient resources, the BS may initiate procedures to move the MS to a different BS. If the service flow is admitted, the BS creates a new SFID and CID, and sends a DSA-RSP containing the admitted QoS parameter set. The MS finalizes the service flow creation by sending a DSA-ACK message.

A BS initiates the creation of a service flow by a sending a DSA-REQ to the MS or SS, as shown in Figure 6.5(b). The MS responds with a DSA-RSP indicating acceptance or rejection of the service flow. The exchanges completes with the BS sending a DSA-ACK message.

Admitted and active service flows can be modified, that is, have their QoS parameters set modified, using a DSC exchange. Similar to the DSA exchange, DSC is a three-way handshaking process. DSC messages can also be used to change the status of a service flow from being active into admitted and to de-admit a service flow. If a DSC message includes values for both admitted and active QoS parameters, the admitted set is checked first. If approved, the active set is checked against the admitted set to ensure that it is a subset. If all checks are successful, the QoS parameters sets in the message become the new admitted and active QoS parameter sets for the service flow. If either of the checks fails, the modification process is considered unsuccessful and the service flow QoS parameters remain unchanged.

The deletion process of any services flow and its associated connections is mandatorily initiated by the BS and optionally by a MS. A BS (MS) wishing to delete a service flow exchanges a three-way handshake DSD message with

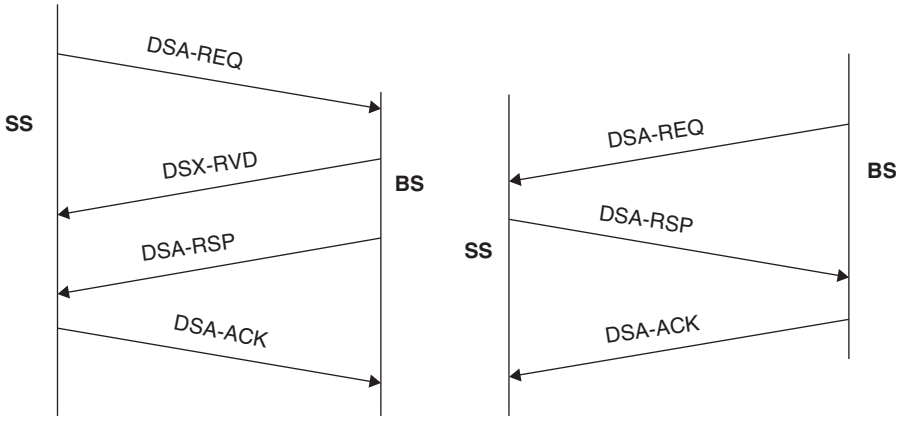


Figure 6.5 Service flow initiation (a) MS-initiated, (b) BS-initiated. Reproduced by permission of © 2009 IEEE.

the MS (BS). When the deletion process is completed, all the relevant allocated resources are released.

6.1.3.2 Bandwidth Request and Grants

To receive allocations on the uplink, an MS must generate a bandwidth request. The requirements of an MS are calculated based on the number of bytes needed to carry its MAC PDUs. The physical layer overhead is excluded from this computation as it depends on the channel conditions. A bandwidth request may be sent either in a dedicated bandwidth request PDU or (optionally) piggybacked using the grant management subheader. A bandwidth request for a connection can also be either incremental or aggregate. For incremental requests, the BS combines the new bandwidth requirements to those of the MS's currently active connection. On the other hand, aggregate requests are treated as a new view of the MS's total bandwidth requirement. Supporting the incremental bandwidth request is optional for SSs but mandatory for BSs, while supporting aggregate BRs is mandatory for both.

An MS requests bandwidth per connection CID, while the BS processes these requests on a per MS basis. In other words, grants are made for an MS as a whole and not for individual CIDs. An MS may receive grants that are less than what it requested and, in these cases, may back off and send a new bandwidth request. The standard mandates that an MS schedules its allocated transmission opportunities for its own connections. However, the standard does not specify the details of such scheduler.

An MS requires allocations to send its bandwidth request. This process is called polling. Bandwidth requests for non-UGS scheduling services can be sent

using either unicast polling, multicast and broadcast polling, contention-based CDMA bandwidth request and the PM bit.

- *Unicast polling*: (applies to SC, OFDM and OFDMA) Each MS is individually polled by the BS. Sufficient resources are allocated for the MS to send its bandwidth request. These resources are indicated in the UL-MAP. Allocations are made per MS and not per connection and are addressed to the MS's Basic CID. An MS must reply to the unicast polling even when it has no data to send. An MS with no current requirements responds with a zero bandwidth request or a dummy PDU with stuffed bytes.
- *Multicast and Broadcast Polling*: (applies to the single carrier and OFDM physical layer) Used when there is insufficient resources to individually poll inactive MSs. The BS allocates transmission opportunities and indicates these allocations in the UL-MAP. Multicast polling is addressed to multicast groups and is associated with a Multicast CID, while broadcast polling is associated with a Broadcast CID. Unlike unicast polling, an MS is not obliged to reply to a multicast/broadcast polling to reserve transmission opportunities. All SSs associated with a broadcast or a multicast polling group can contend for the shared allocated bandwidth to send their BRs. Collisions are resolved by using a truncated binary exponential backoff algorithm. An MS assumes collision if it did not receive grants in the subsequent UL-MAP messages received before the contention-based reservation timer expires.
- *Contention-based CDMA bandwidth request*: (applies to OFDMA) An MS can use the ranging subchannel and contend using a bandwidth request ranging code, as discussed in Chapter 5. When the BS detects the ranging code, it responds with a CDMA Allocation IE to specify the transmit region and transmitted ranging code over a Broadcast CID. The MS may use the transmission region to send a bandwidth request or data. If the MS does not receive a reply from the BS within a predefined time period, it assumes collision and performs a contention resolution procedure.
- *PM bit*: An MS with a currently active UGS connection can indicate that it needs to be polled for non-UGS connections through setting the PM bit in a PDU's GMSH within its UGS connection. Once the BS detects that a PM bit is set, it initiates unicast polling for the requesting MS in the subsequent transmission opportunities. To minimize the risk of the BS missing the PM bit, the MS may set the bit in all UGS PDUs in an uplink scheduling interval.

6.1.4 Bandwidth Allocation and Traffic Handling

The standard does not specify any requirements for traffic policing and shaping within an IEEE 802.16 network. For scheduling, a scheduling algorithm for the UGS service (CBR) traffic called persistent scheduling is defined for OFDMA. Persistent Scheduling is a technique used to reduce MAP overhead

for connections with periodic and fixed payload-size traffic. UGS resources are persistently allocated by the BS. The BS transmits a Persistent HARQ DL MAP IE for downlink allocations and a Persistent HARQ UL MAP IE for uplink allocations. The persistently allocated resources are maintained during the life of a UGS service and are released at the termination of a UGS service. No specific scheduler is defined for other scheduling services, whether for downlink or uplink.

6.1.4.1 Automatic Repeat Request and Hybrid Automatic Repeat Request

The standard's support for error control mechanisms is optional. The supported mechanisms are ARQ and HARQ, with the latter additionally featuring error correction. Both utilize a mixture of retransmissions and timeouts.

When a PDU is transmitted in ARQ, the transmitter starts a timer and waits for an acknowledgment from the receiver that reception was successful. However, if the timer expires before the acknowledgment is received, the PDU is considered lost, and a retransmission is scheduled through the ARQ process. Enabling ARQ over a connection automatically enables it for all the PDUs of this connection. The ARQ process partitions the MAC SDU into blocks whose length is specified by the ARQ BLOCK SIZE parameter. This latter is set during the connection establishment. Each block has a Block Sequence Number (BSN) that is included in the fragmentation and packing subheader of the ARQ-enabled connections. A receiver sends an ACK or negative ACK (NACK) feedback as a response of a received PDU. The feedback can be sent as a stand-alone MAC PDU over the basic management connection or be piggybacked to a data MAC PDU. Three ARQ modes are supported: Stop and Wait, Go back N and Selective Repeat.

HARQ[1] employs error correction in addition to error detection. Unlike ARQ, where PDUs are individually coded, HARQ utilizes a connection's earlier PDUs to decode the PDU most recently received. This feature effectively reduces error probability and enhances throughput through reducing the number of required retransmission per PDU. However, this is achieved at the cost of increased receiver complexity.

Based on the nature of the retransmitted replica, the standard defined two types of HARQ. These are:

- *Chase combining*: Each retransmission is identical to the original transmission. Hence, there is need to identify each retransmission, and decoding the most recently received PDU can be made by combining it with all previously received transmissions.
- *Incremental redundancy*: This is an improvement on chase combining whereby different versions of the first PDU transmissions are made. The error correction and detection bits patterns are different from the earlier transmissions, while the information bits remain the same. Consequently, the receiver gains additional

[1] This instance of HARQ is more accurately described as HARQ with soft combining.

knowledge (compared to gaining additional energy in the chase combining) which helps it recover the erroneous bits.

6.2 Quality of Service in IEEE 802.16j

To facilitate MR, the IEEE 802.16j amendment included additional features and specification. However, it does not specify additional QoS parameters to the ones mentioned above. Consequently, we shall start directly at the classification process.

6.2.1 Classification

A MAC PDU crosses a tunnel connection, where it is classified based on the ID of this connection. If it is not addressed to a tunnel connection, it is classified based on its own CID. Note that the connection may be a member of the tunnel. In such a case, the connection PDUs are addressed to the connection's destination and not the tunnel's.

The amendment also defines routing paths between the MR-BS and newly attached MSs or RSs. After it discovers any topology changes due to a node joining or leaving the network, for example, due to mobility, the MR-BS removes an old path or establishes a new one and informs all RSs along that path. A new connection, established along this path, is bound to it using a Path-ID and the CID. Moreover, it can be either, an individual connection or a tunnel connection.

6.2.2 Signaling Bandwidth Requests and Grants

Connections with similar QoS parameters and the same ingress-egress pair join the same connection tunnel. An MR-BS determines the service flow parameters associated with a tunnel and distributes these parameters to all RSs along the tunnel's path. When new connections for an MS are added to or removed from a tunnel, an MR-BS modifies the tunnel's service flow parameters using a three-way handshake and the dynamic service flow addition, creation and deletion procedures. The same takes place when the QoS requirements of the individual connection for MS are changed provided that this change modifies the tunnel's QoS requirements. The following details the different IEEE 802.16j procedures for service flow creation, change and deletion; and path establishment and removal, in addition to the bandwidth request and grant procedures.

6.2.2.1 Service Flow Creation, Change and Deletion

If an MS initiates a request for service flow creation with a scheduling RS, the BS will seek the admission decision from all RSs along the path before it accepts or rejects the request. If the service flow requested is mapped to an existing tunnel associated with service flow parameters, and if the service flow request will result

in changing these parameters, the MR-BS will send a Dynamic Service Change Request (DSC-REQ) to all RSs along the path between the MR-BS and the MS over the primary management connection. Once an RS receives the request and decides that it can support the requested QoS parameters, it forwards this request to its subordinate RSs. If the RS cannot support the request, it sends a rejection Dynamic Service Change Response (DSC-RSP) that may contain information about the RS's own QoS parameters.

To ensure that the DSC-REQ messages follow the same path as the MAC PDU associated with the tunnel information, the MR-BS may include a path_ID TLV explicitly identifying the route of this request in the DSC-REQ. All intermediate RSs use this path ID to route the DSC-REQ message. This method is called explicit path management. The MR-BS may follow another procedure to define the path of the DSC-REQ message (and consequently the users' data path) called the embedded path management. In this procedure, a Path Info TLV is included in the DSC-REQ to inform each RS of the primary CID of its subordinate RS to identify the next hop of the DSC-REQ. If this information is not included, the intermediate RS determines the next hop by checking the transport CID included in the services flow parameters in the DSC-REQ.

If all RSs along the path can support the QoS requested parameters, the MR-BS receives DSA/DSC-RSP from the access RS within a time limited by timer T59. The MR-BS then sends DSA/DSC-RSP to the requesting MS, and a DSA/DSC-ACK with the admitted service flow parameter to all the RSs on the path using the same route used to send the corresponding DSA/DSC-REQ. This completes the three-way handshake required for service flow addition.

A service flow addition may be initiated by the MR-BS to an MS to set up an individual service flow or to an RS to set up a tunnel service flow. In these cases, before the MR-BS creates a DSA-REQ to an MS, it would send a DSA/DSC-RSP to all RSs on the path to verify resource availability. The procedures of sending and processing the DSA/DSC-REQ and DSA/DSC-RSP are the same as explained above.

The procedure for changing a service flow is similar to the procedure of adding new service flow. The only difference is that instead of sending a DSA-REQ from the MR-BS or the MS, the message sent is a DSC-REQ and the REQ-RSP-ACK procedure is performed.

Finally, to delete a service flow in a network with scheduling RSs, the MS or the BS may initiate the procedure by sending a DSD-REQ for an existing service flow. For an MR-BS initiated request, the MR-BS sends a DSD-REQ to all the RSs on the path if the service flow is not mapped onto a tunnel, or if the service flow is mapped onto a single transport tunnel of an access RS and the tunnel has no service parameter associated. Otherwise, if a change of the tunnel's service flow parameters is required to delete this service flow, the MR-BS sends a DSC-REQ to all RSs on the path. For an MS initiated request, once an MR-BS receives a request from the MS it will send DSD/DSC-REQ message as explained above. The deletion of a service flow or a tunnel procedure follows the REQ-RSP-ACK three-way-handshake.

6.2.2.2 Path Establishment and Removal

An MR-BS calculates a path for the uplink and downlink traffic between itself and an access RS over a topology tree. The topology tree is calculated centrally at the MR-BS given the topology information obtained from the topology discovery or update process. The path calculation is constrained by considerations such as the tree topology architecture, the availability of radio resource, quality of the link and load condition of the RSs. Information about a certain path is distributed to RSs in DSA-REQ messages as discussed earlier. MR-BS uses either explicit path management or embedded path management for disseminating the path information. However, algorithms for tree construction and path calculation are not defined in the standard.

In Explicit path management, the path information and a uniquely assigned path ID are included in DSA-REQ message. The CIDs to be routed on this path and their associated service flow parameters may also be included for path/CID binding operation. The RSs receiving the DSA-REQ message determine the next hop of the message among its neighboring RSs by discerning the primary management CID of the subordinate RS from the DSA-REQ message. This process is repeated at every intermediate RS until the access RS is reached. The access RS then sends a DSA/DSD-RSP directly back to MR-BS. If an intermediate RS fails to process the request, it sends a DSA-RSP directly to the MR-BS with the associated confirmation code. If the MR-BS decides to remove an existing path, it sends a DSD-REQ message with the path ID using explicit path management. The RSs receiving the DSD-REQ message shall remove all the information related to that path using the same procedure discussed above for determining the next hop.

In embedded path management, the MR-BS systematically assigns set of CIDs to each of its subordinate stations, and assigns a subset of RS's allocated CIDs set to all subordinate RSs of any of that RS. Using this systematic CID structure helps RSs find routing paths without storing all CIDs of subordinate RSs in the routing table. If an MR-BS used the embedded path management for DSA/DSD-REQ message, the message includes a Path Info TLV to inform each RS of the primary CID of its subordinate RS, which identifies the next hop of the message.

6.2.2.3 Bandwidth Request and Grants

The bandwidth request and grants procedures depend on the type of relaying, that is, whether transparent or non-transparent, and the scheduling employed, that is, whether centralized or distributed. In distributed scheduling, a non-transparent scheduling station directly handles the bandwidth request it receives from its subordinate RS. A subordinate RS may send a bandwidth request using the MAC signaling header, the grant management subheader or the CDMA bandwidth request code. The bandwidth request may be a standalone request or it may be piggybacked to a relay data MAC PDU.

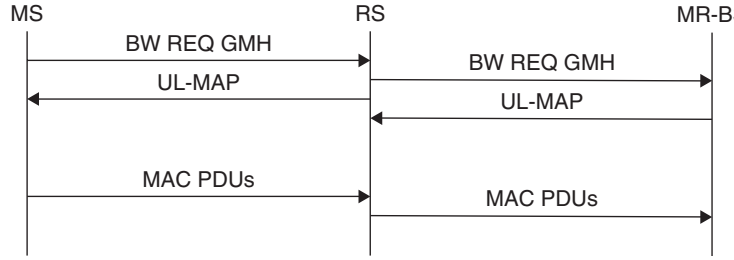


Figure 6.6 An Example of a bandwidth request forwarded by the superordinate as soon as it is received from the subordinate. Reproduced by permission of © 2009 IEEE.

As in the PMP mode, a bandwidth request in MMR may be incremental or aggregated. Supporting incremental BRs is only mandatory for non-transparent scheduling RSs. Once an RS receives BRs from its subordinates stations it may combine the amount of bandwidth requests received from its subordinates to its own bandwidth needs to generate one bandwidth request header for each QoS class. Alternatively, an RS may transmit a bandwidth request from one of its subordinates as soon as it is received. This procedure is shown in Figure 6.6.

In centralized scheduling, the MR-BS determines all bandwidth allocations for every station in the MR cell, and includes information about these allocations in its MAP messages. A non-transparent superordinate station does not combine its subordinate BRs. Similarly, the superordinate station is only required to forward the subordinate request uplink to the MR-BS.

In distributed scheduling, a superordinate station informs its subordinate scheduling RSs about their bandwidth allocation ahead of time using the RS-SCH message. The RS-SCH message includes when the bandwidth is allocated in number of frames, the size of allocation, and the intended CID. The actual bandwidth grant is signaled to the subordinate by the non-transparent scheduling RS in the relay UL-MAP message. Once an RS receives an RS-SCH management message from its superordinate station (MR-BS or RS), it shall look up the "next hop" of the given CID. Based on this scheduling information and the next of the CID, the RS can determine the appropriate bandwidth allocations and associated RS uplink allocation frame offset on the uplinks it oversees. The RS sends its own RS-SCH management messages to its subordinate RSs to inform them of the bandwidth allocation decisions it makes. This process is repeated until the RS-SCH reaches the access RS.

In centralized scheduling, an MR-BS allocates uplink and downlink bandwidth for each station in the MR-Cell and over all links that make up the path between that station and the MR-BS. For successful bandwidth allocations and continuous forwarding of a MAC PDU on consecutive links along a path, the MR-BS is required to create the bandwidth scheduling taking into account processing delay at each RS, the multihop frame structure of the MR cell, and link qualities at each RS along the path. To facilitate accounting for delay at the MR-BS, access RS and

intermediate RSs, each station informs the MR-BS of its minimum forwarding delay capability using the SS Basic Capability Request (SBC-REQ) message during the RS's network entry process.

The amendment describes additional functionality for supporting polling over the multihop architecture. An MR-BS or RS informs a subordinate RS of upcoming polling using an RS Scheduling (RS-SCH) management message. In centralized scheduling, only the MR-BS performs the polling process. When an MR-BS polls an MS/RS, it schedules the polling process so that each intermediate RS along the path to the target MS/RS is polled sequentially.

6.2.3 Bandwidth Allocation and Traffic Handling

6.2.3.1 Scheduling

While the amendment defines the required changes in signaling messages, bandwidth request and grant mechanisms, and ARQ/HARQ, it does not define specific scheduling algorithms for either the centralized or distributed modes.

6.2.3.2 ARQ/HARQ

The amendment supports three ARQ types: end-to-end, two-link, and hop-by-hop. The type of ARQ operation mode is negotiated and agreed upon during the RS's network entry.

In end-to-end ARQ, the process is maintained by the MR-BS and the MS, while the intermediate RSs merely forwarding the exchanged packets. On the other hand, two-link ARQ, as the name suggests, maintains two ARQ processes: one between the MS and the access RS while the other is between the access RS and the MR-BS. When there are intermediate RSs, their role is just to forward and feedback the ARQ-enabled PDUs between the access RS and the MR-BS. Finally, in hop-by-hop ARQ, a separate ARQ process is maintained between each two consecutive elements along the path between the MR-BS and the MS.

An MR-BS in two-link ARQ sends an ARQ-enabled PDU to the access RS and awaits feedback from the access RS. If the PDU is in error over the relay link, the access RS sends NACK and the MR-BS schedules a retransmission. If the MR-BS receives a R-ACK from the RS, it awaits the ACK from the MS. If the PDU is in error over the access link, the access RS schedules a retransmission to the MS until either an ACK is received or until a timer (defined by ARQ_BLOCK_LIFETIME) expires. The MR-BS and the RS retry timers in MR-BS are independent. Similar procedures apply for the uplink.

For a successful downlink transmission in hop-by-hop ARQ, the MR-BS receives a R-ACK from its subordinate RS and an ACK from the MS relayed on hop-by-hop basis along the path between the MR-BS and the MS. For a corrupted downlink transmission the subordinate RS does not relay NACK to its superordinate. The subordinate RS, however, keeps scheduling the PDU retransmission until an ACK is received or the retransmission timer expires.

The amendment also describes HARQ for both the centralized and distributed scheduling modes. In centralized scheduling, the MR-BS schedules initial transmission of an HARQ-enabled PDU on all the links along the path between the MR-BS and the MS. Failure of transmission at any hop along the path is signaled to the MR-BS and relayed on the uplink ACK/NACK channel. In this case, the MR-BS signals the HARQ burst allocations to the failed link and all the subsequent links in the RS_HARQ_DL_MAP, in case of downlink transmission, and RS_HARQ_UL_MAP, in case of uplink transmission. These allocations are made for the stations to send the uplink ACK/NACK. A BS identifies the link in failure from the code error and the CID or the Reduced CID (RCID) encoded in the uplink ACK/NACK message.

The amendment also supports HARQ operation for the group RS. In this case, a shared ACK/NACK channel is allocated for the whole group. Scheduling and shared ACK/NACK channel allocation is made by the group's superordinate (whether MR-BS or ntRS) HARQ operation in RS groups can be performed in any of the three ARQ modes.

6.3 QoS in IEEE 802.16m

6.3.1 QoS Parameters

The IEEE 802.16m amendment does not define any additional QoS parameters other than the QoS parameters defined in IEEE 802.16-2009.

6.3.2 Classification

The IEEE 802.16m amendment defines a new service flow type called the emergency service flow. These are given priority in admission control over regular service flows. Default service flow parameters are defined for emergency service flow. The ABS grants resources in response to an emergency service notification from the AMS without going through the complete service flow setup procedure. Each PDU crossing an IEEE 802.16m network is associated with a unidirectional flow of packets possessing a specific QoS requirement with a service flow. Each service flow is mapped to one transport connection that is defined by one FID and one STID. The scheduling services (UGS, rtPS, nrtPS, BS and ExrtPS) of the WirelessMAN OFDMA reference system are supported in IEEE 802.16m. The IEEE 802.16m amendment also provides a specific scheduling service to support real time non-periodic applications such as on-line gaming.

6.3.3 Bandwidth Request and Grant

In IEEE 802.16m, BRs are transmitted through either indicators or messages. bandwidth request messages can include information about the status of queued

traffic at the AMS such as buffer size and quality of service. IEEE 802.16m also supports modifying QoS parameters for active flows. The AMS and ABS negotiate the supported QoS parameter sets during service flow setup procedure. When QoS requirement/traffic characteristics for uplink traffic changes, the ABS may autonomously switch the service flow's QoS parameters such as grant/polling interval or grant size based on predefined rules. In addition, an AMS may request the ABS to switch a service flow's QoS parameter set with explicit signaling to allocate resources according to a new set.

6.3.4 Bandwidth Allocation and Traffic Handling

The IEEE 802.16m amendment does not define specific scheduling algorithms for either single or multihop Advanced networks. The amendment, however, states that an ARS may operate in either a centralized or distributed mode. When an ABS is configured to operate in centralized scheduling, the ABS schedules all radio resources in its cell. In distributed scheduling, each station (ABS or ARS) schedules the radio resources on its subordinate link within the radio resources assigned by the ABS.

The described ARQ mechanism is similar to the operation of that in IEEE 802.16-2009. However, HARQ is mandatory in IEEE 802.16m, and is of two types: chase combining and incremental redundancy. The HARQ is an N channel Stop and Wait mechanism that uses adaptive asynchronous HARQ in the downlink and adaptive synchronous HARQ in the uplink.

In adaptive asynchronous HARQ, the resource allocation and transmission format for the HARQ retransmissions may be different from the initial transmission. Any retransmission is scheduled by the ABS and information about allocations for this retransmission is signaled to the AMS using the control message Advanced allocation map (A-MAP). Once an AMS receives the A-MAP, it recognizes and identifies the downlink burst destined to it from the ABS. If the AMS decodes this burst correctly, it responds to the ABS with an ACK. Otherwise, the AMS sends a NACK to the ABS and a retransmission of the failed data has to be scheduled by the ABS within the data maximum retransmission delay bound. An HARQ burst is discarded if a maximum number of retransmissions is reached. For constant bit rate traffic, which has a persistent allocation on the initial transmissions, HARQ retransmissions are supported in a non-persistent manner, that is, resources are allocated dynamically for HARQ retransmissions.

For synchronous HARQ, which is utilized in the uplink, resource allocation for the retransmissions can be fixed or adaptive. However, the default operation mode of HARQ in the uplink is non-adaptive, that is, the parameters and the resources for the retransmission are known *a priori*. Using signaling, the ABS can enable an adaptive uplink HARQ mode. When enabled, the parameters of the retransmission are signaled explicitly. An AMS that is allocated an uplink bandwidth is informed about its allocation using the uplink A-MAP. If the ABS's

decoding is successful, the ABS sends ACK to the AMS. Otherwise, the ABS will send a NACK to the AMS. Upon receiving the NACK, the AMS triggers the retransmission procedure. If during retransmission the AMS does not receive a uplink A-MAP for the HARQ data burst in failure, the AMS transmits the failed PDU through the resources assigned at the latest PDU transmission opportunity with the same ACID. If the uplink A-MAP is assigned, the AMS performs the HARQ retransmission as instructed in this uplink A-MAP.

7

Mobility Management

IEEE 802.16 provides efficient mobility support for higher management structures, for example, Mobile IP. As is typical of cellular systems, BSs in IEEE 802.16-2009 oversee much of the mobility management signaling. In relay systems, some responsibilities can be assigned to scheduling ntRSs in order to maintain their autonomy and network efficiency. While inter-RAT handovers are accommodated in IEEE 802.16-2009, the IEEE 802.16m amendment expands this accommodation to allow greater flexibility in management and operation.

This chapter is organized as follows. Section 7.1 discusses mobility management in IEEE 802.16-2009, while Sections 7.2 and 7.3 discuss mobility managements in the amendments IEEE 802.16j-2009 and IEEE 802.16m, respectively. A particular emphasis is made in Section 7.1 in describing the handover process, which underwent certain optimizations in IEEE 802.16m. Section 7.2 elaborates on the added considerations made when relay networks are employed, including when a relay station itself is mobile. Section 7.3, on the other, describes both purely Advanced and mixed Advanced-Legacy IEEE 802.16 handovers will be carried. The section also describes particulars of handovers of inter-RAT and femtocell mobility.

7.1 Mobility Management in IEEE 802.16-2009

Figure 7.1 shows the procedures involved in initiating and carrying out a handover. The standard does not specify how the handover decision should be made, nor does it mandate whether the decision should be made by the network or the MS. The standard, however, provides means for information acquisition by both the BS and the MS to make efficient decisions. The information acquired generally specifies the quality of the signal received by the MS from the various BS, but also information on the readiness of these BSs to support the MS's requirements.

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

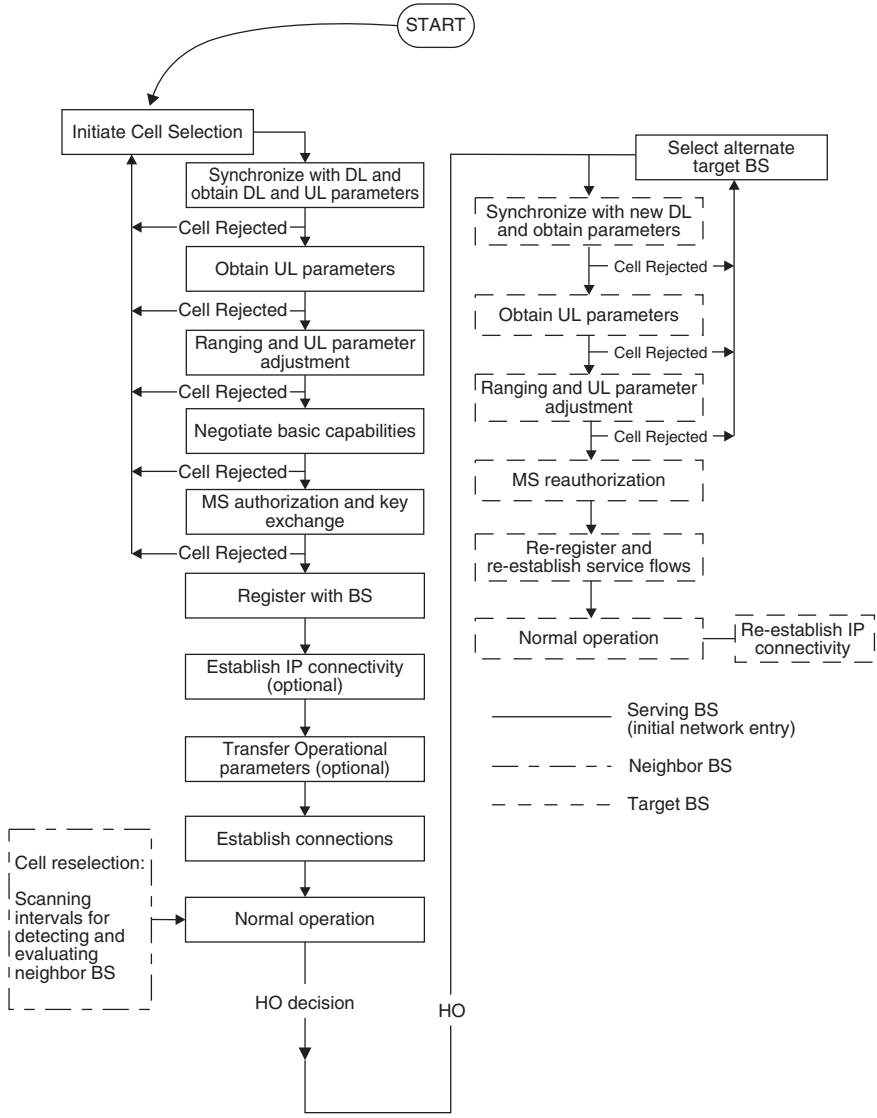


Figure 7.1 Flow chart for handover process. Reproduced by permission of © 2009 IEEE.

The procedures in the figure are almost identical to those of network entry and initialization, which were described previously in Chapter 5. To enhance service delivery for an already active call, the standard defined certain optimizations so as to accelerate the handover process. These optimizations are described on the next page. In addition, certain enhancements in the standard, such as seamless handover and macro-diversity handovers, will be outlined.

7.1.1 Acquiring Network Topology

A network topology means different things for the BS and the MS. A BS requires an understanding of the capabilities of its neighboring BS, and requires active communication links to these BS for various mobility management objectives, including handover and resource allocations. Meanwhile, an MS must have a way of measuring the signal strengths and understanding the capabilities of the BSs it recognizes as traverses the network.

A BS understands the status of its neighboring BSs through the network's backbone. For an MS, information on network topology can be acquired in two manners: either through topology advertisement from the serving BS, or through the MS scanning for neighbor BSs. The topology advertisement, sent through the neighbor advertisement message (MOB_NBR-ADV), voids the need for the MS to scan the DCD/UCD of neighboring BSs. Each MS maintains what is called an association table that maintains a record of the BSs to which the MS has been or may be associated. The table is limited by a parameter specifying the maximum number of neighbors, $N_{MS\_max\_neighbors}$. A BS may send information for more than $N_{MS\_max\_neighbors}$ neighboring BSs. However, an MS is only required to support at $N_{MS\_max\_neighbors}$ in its association table.

To perform scanning of neighboring BSs, an MS needs to be allocated time intervals from its serving BS. An MS therefore needs to send a scanning interval allocation request (MOB_SCN-REQ), through which an MS may specify desired scanning intervals with interleaving intervals of operation. The final decision for selecting scanning intervals and their durations, however, is completely left to the BS. A BS receiving a MOB_SCN-REQ responds with MOB_SCN-RSP either granting scanning intervals and durations, or denying the sensing request. In a MOB_SCN-RSP, a BS may also recommend certain neighboring BSs for association. A BS may also send an unsolicited MOB_SCN-RSP whereby the MS would respond with a report of its measurements for the indicated neighboring BSs, if applicable. It is possible that an MS receives several responses, called scanning interval allocation response or MOB_SCN-RSP, to its scanning request, which could happen if multiple BSs respond to the MS's request, or if the MS's serving BS has a short timeout clock. In such cases, the MS should only consider to the most recent it receives. Whether or not a BS recommends neighboring BSs for association, an MS can perform any scanning or association activities in the scanning intervals it is allocated.

7.1.2 Association Procedures

The standard defines association as "an optional initial ranging procedure occurring during scanning interval with respect to one of the neighbor BS." Its role is to enhance the handover process through gathering information that is useful for selecting the target BS and that accelerates the MS's ranging process during handover. As aforementioned, an MS maintains an association table for the various BSs that it is informed about or becomes aware of as it traverses

the network. To maintain a fresh list, each BSID entry in the association table expires after a set amount of time. This expiry can be either indicated by the BS, or set by the network. If indicated in initial ranging, a serving BS can be included in the table.

For an MS, association can be directed by a BS recommending neighboring BSs through MOB_SCN-RSP. If the MS supports directed association, it shall scan the recommended BSs. Such support is indicated during network entry when negotiation basic capabilities.

There are three levels of association.

- *Association Level 0*: Scan/Association without coordination;
- *Association Level 1*: Association with coordination; and
- *Association Level 2*: Network assisted association reporting.

If Level 0 is chosen by the network, only the scanning intervals are coordinated between the serving BS and the MS. A target BS would not be aware of an MS possible handover to its coverage. The MS would contend as if performing initial ranging.

If the MS requests (through MOB_SCN-REQ) or the BS arranges for a Level 1 association, the BS would provide scanning intervals to the MS and coordinate handover with neighboring BSs. Through an unsolicited MOB_SCN-RSP, the serving BS may indicate possible alternatives to the MS. A neighboring BS would provide a rendezvous time, a unique code number and a transmission opportunity. A rendezvous time is the identification of the frame with the transmission opportunity in which the MS would send the unique code number. Assigning a unique code and transmission opportunity is called dedicated ranging. A form of coordination may be exercised between BSs such that no or minimum collision can result from codes utilized in handover procedures. An MS is expected to synchronize immediately at the first frame following the rendezvous time, and be able to extract information from the UL-MAP to distinguish the transmission opportunity. If synchronization fails, the MS aborts the Level 1 association. If the MS is still interested in establishing the connection, the MS may perform a Level 0 association afterwards.

In the network assisted association, Level 2, an MS is only required to send the CDMA ranging to the neighboring BS at the appropriate time. The serving BS would relay ranging and other information from the neighboring BS through a single association report message (MOB_ASC_REPORT). The target BS would expect the MS's CDMA code during the ranging period indicated, whether or not it is a dedicated ranging period.

7.1.3 The Handover Process

Procedures for cell reselection and handover decision and initiation need not to be associated to each other. An MS may consider reselection proactively to achieve certain operational objectives, for example, an MS may persistently seek the BS

with a relatively much higher signal strength or with better indications of QoS support. Handover decisions may originate at either an MS or the serving BS. If an MS decides to handover, it sends an MS handover request (MOB_MSHO-REQ), while if a BS is requesting an MS to handover; it sends a BS handover request (MOB_BSHO-REQ). For any ongoing handover process, both the MS and BS will ignore any handover initiation requests unless the ongoing handover has been acknowledged to be either successfully terminated or cancelled.

An MS needs to synchronize and range with the target BS prior to resuming regular operations. Depending on how association is made, synchronization and ranging can be optimized to accelerate the handover process. In particular, procedures for negotiating basic capabilities, authentication, registration and adjustments can all be skipped during handover procedure. As such information would have been relayed in initial ranging, it is possible to establish a level of collaboration between the BS in IEEE 802.16 such that these information can be relayed from the serving BS to the target BS.

If an MS acquires a pre-allocated Basic CID prior to a seamless handover, it will be able to derive the primary management and transport CID autonomously from the pre-allocated basic CID. This usually takes place if a seamless handover can be supported at the serving BS, the target BS and the MS. If a dedicated allocation is made for the MS to send in a RNG-REQ directly, the CDMA ranging can be bypassed. Such a dedicated allocation would be part of what is called fast ranging – an expedited ranging procedure that is viable through the serving BS negotiating entry parameters with the target BS. The target BS would indicate further information using a Fast_Ranging_IE information element.

Options within the MOB_BSHO-REQ and MOB_MSHO-REQ offer great flexibility in terms of how a handover can proceed. A BS, for example, can specify the target BS to which the MS should attempt to handover. As an alternative, the BS may select a group of neighboring BS, of which the MS can attempt handover to one or more. In such a case, the MS does not need to notify the serving BS about the chosen BS. It is also possible that an MS would ignore BS's recommended set of neighboring BSs. An MS can also explicitly reject a handover recommendation if it is unable to successfully perform the process. In such a case, a BS may reconsider its recommendation. In some instance, a BS can force the MS to perform handover regardless of the MS's considerations of choice. Note that a serving BS may coordinate with neighboring BSs, informing more than one BS of the MS's intent to handover. Information, as will be described below, can also be relayed to expedite the handover procedure.

An MS committed to a handover will terminate with the serving BS through indicating the handover type in a handover indicator message (MOB_HO-IND). In such instances, a BS may retain certain information about the MS that would expedite the MS handover to the target BS and would indicate this to the MS in the MOB_BSHO-RSP message. A MS can cancel an ongoing handover procedure through the serving BS by either resuming regular operation, for example, sending a bandwidth request, or by explicitly cancelling the handover in a MOB_HO-IND message.

A handover drop occurs when an MS loses communication with the serving BS prior to completing the handover procedure, including termination with the serving BS. Both an MS and a BS can detect a drop, for example, if the number of RNG-REQ retries limit has been exceeded. An MS detecting a drop can attempt a handover entry to either a target BS or the serving BS. If the serving BS has already discarded the MS's context, a full network reentry with possible handover optimization needs to be performed.

If both the serving and the target BSs support continuity of downlink transmissions, for example, maintainable state for ARQ connections through the handover, it is possible that the "in flight" information can be transferred from the serving BS to the target BS to maintain continuous delivery at the MS.

The standard describes a range of possible optimizations whereby an MS's context can be shared between the serving BS and the target BS. An MS's static context refers to parameters acquired and configured in network entry and initialization, and that may have been adjusted later on during an MS's connection lifetime, in addition to all service flow encodings. An MS's dynamic context, however, refers to the state of counters, timers, state machine status, and data buffer contents.

There are several levels of handover optimization that can be exercised in the network. At one end, no optimizations can be exercised, and in such a setting an MS always has to perform a full network entry with or without a traffic IP address refresh. At the other end, there is the fully optimized handover whereby both static and dynamic contexts are handed from the MS's serving BS to the target BS during handover. In between the two ends, there are further two options whereby full optimization can be done with, for example, Traffic Encryption Key (TEK) updates, and a partially optimized handover whereby only static context is moved.

7.1.4 Optional Handover Modes

The standard defines two optional handover modes, both of which are based on diversity communications. The two modes, called Macro-Diversity Handover (MDHO) and Fast BS Switching (FBSS), essentially rely on maintaining diversity set detailing the set of BSs with which an MS can establish connections. The difference between the two handover types is that in MDHO the MS communicates with all BSs in the diversity set, while in FBSS the MS only communicates with an anchor BS. A critical advantage of FBSS is that an MS does not perform a full handover process; rather, it merely indicates a change of anchor in its diversity set. In other words, an MS "serving" entity becomes the diversity set, and not just one serving base station. As long the MS remains within the same diversity set, it is not required to perform a full handover procedure. Both MDHO and FBSS can be viewed as soft handovers, compared to the mandatory hard handover described above.

BSs involved in MDHO or FBSS, in addition to the MS, must support the type of handover mode utilized. For both BSs and MSs, the support of either mode

is optional. An MS is to follow the type of handover dictated by the BS either in its response to a MOB_MSHO-REQ, that is, a MOB_BSHO-RSP, or in its handover initiation, that is, a MOB_BSHO-REQ. As in a regular handover, both the BS and the MS would ignore any handover initiation requests if there is a currently active one.

The MS selects and updates BSs for its diversity set through scanning, and shall report this diversity set to the serving or anchor BS, depending on the handover mode employed. An MS is also required to continuously monitor the signal strength of the BSs included in this set, and select one BS as its anchor. Meanwhile, an MS may consider the anchor's MOB_NBR-ADV, previously performed signal strength measurement, propagation delay measurement, scanning, ranging and association activity. The selection should be reported through either the CQICH or the MOB_MSHO-REQ. A regular handover can be considered a special case of either MDHO or FBSS whereby the diversity set includes a single BS that is also the anchor BS.

BSs supporting either diversity handovers would include the H_Add and H_Delete thresholds in their DCD messages. These thresholds can be used by an MDHO or FBSS capable MS in determining whether a BS should be included or deleted from the diversity set maintained by the MS. If the mean CINR of an active BS in the current diversity set is less than the H_Delete threshold, the MS may request that this BS be dropped from this diversity set. Similarly, if the mean CINR is greater than the H_Add threshold, the MS may request that the respective BS be added. The MS update of its diversity set is recommended but not mandated. In fact, an MS's diversity set is only required to be a subset of those listed in a BS's MOB_BSHO-RSP or MOB_BSHO-REQ. An MS may reject a recommended diversity set any recommend a preferred one to be included.

There are two ways in which an MS can gather control information under MDHO. In the first one, the MS observes control information from all BSs in the diversity set while in the second one, the MS only observes control information from the anchor BS. In the latter case, the anchor BS may include burst allocation information for the non-anchor BS. Under FBSS, an MS only observes the anchor BS for control information.

An MDHO begins with a decision for an MS, made by either the MS or the BS, to begin a simultaneous exchange of messages and traffic with multiple BSs. For the downlink, two or more BSs would provide synchronized transmission to the MS; while for the uplink, transmission from the MS would be received by multiple BSs. For FBSS, a handover comprises an update of the anchor BS.

The following conditions are shared by both handover types:

- The involved BSs are synchronized based on a common time source.
- The frame sent by the involved BSs at a given frame time arrive at the MS within the same prefix interval.
- The involved BSs have a synchronized frame structure.
- The involved BSs have the same frequency assignment.
- The involved BSs required to share or transfer the MS's MAC context.

Similarly, the following conditions pertain only to MDHO handovers:

- The involved BSs would use the same set of CIDs for the connections that are established with the MS.
- The same MAC/PHY PDUs shall be sent to the MS by all the involved BSs.

There are two manners in which the anchor BS can be updated in MDHO employing anchors or FBSS. The first relies on the use of HO messages, whereby handover is achieved after deciding on a preferred anchor and a switching by either the MS (through a MOB_MSHO-REQ and MOB_BSHO-RSP exchange) or the BS (through a MOB_BSHO-REQ). The second manner utilizes the fast-feedback whereby the MS transmits fast anchor BS selection information to the current BS selection. The standard describes the required signaling between the MS and both the old and the new anchors in order to achieve a stable transfer. In both update methods, network entry procedures are not required if the new anchor BS is within the MS's diversity set. The standard also describes procedures for MS-assisted coordination of downlink transmission when an MS performs an anchor BS update, but only under FBSS.

7.2 Mobility Management in IEEE 802.16j-2009

The IEEE 802.16j amendment describes how the MR-BSs and the RSs should behave during MS mobility between MR-BSs and RSs, and between different RS. The amendment also details the signaling required for RS mobility. In both cases, the MR-BS maintains substantial control of MS handover, even when scheduling ntRSs are involved.

7.2.1 MR-BS and RS Behavior during MS Handover

Depending on the type of the MS's target superordinate station, the MR-BS ensures sufficient resources are provided so that signaling between the MR-BS and the MS can be delivered. If the superordinate is an ntRS with centralized scheduling, MR-BS inserts a Fast Ranging IE in the UL-MAP to be broadcast on the access link and provides sufficient bandwidth on the relay link for forwarding the RNG-REQ. If the superordinate station is a scheduling ntRS, it is instructed by the MR-BS to send the Fast Ranging IE. For tRSs, the MR-BS inserts the Fast Ranging IE in the UL-MAP and provides sufficient bandwidth in the tRS's uplink; the tRS, in turn, would forward the RNG-REQ on the access link.

For topology advertisement, each ntRS may advertise differently from the MR-BS's own MOB_NBR-ADV. Under centralized scheduling, the MR-BS must allocate bandwidth for its advertisement; under distributed scheduling, the RSs are autonomous in their allocation.

MR-BS controls the scanning procedure. Scheduling ntRSs coordinate with MR-BS to schedule scanning, and may terminate scanning procedures if they

see fit. For ntRS with unique BSIDs and centralized scheduling, RSs not involved in the handover are notified to ignore handover communication. This applies to RSs within and outside the MR-cell. Scheduling ntRS act similarly but notify the MR-BS of the association. An MR-BS in turn, confirms the association if it sees fit. When an MS perform neighbor scanning Level 0 or Level 1, access stations perform the same tasks as those for contention based initial ranging. RSs may report the observed link quality to the MR-BS. This applies for both tRS and ntRS. Neighboring ntRSs shall inform the MR-BS through a RNG-RSP.

An MS handover requires updating the routing information. A serving MR-BS sends out an MS_INFO-Del to old RS when the latter is no longer supposed to maintain the MS's information. The RS must confirm deletion. Similarly, when a target cell is informed that the MS has attached to a different RS or MR-BS, that is, a drop has occurred, it notifies the old station to delete the MS's context.

For handover optimization, context transfer can be made either by the serving or the target access station. In either case, if the requesting station is an RS, it shall not include information about its respective MR-BS; rather, the MR-BS will augment its own information to the context message.

7.2.2 Mobile RS Handover

When a mobile RS (MRS) is handed over from one access station to the other, it follows procedures similar to those of a regular MS. As depicted in Figure 7.2, however, additional steps are required in order to maintain connectivity for both the MRS and the MSs it oversees.

These additional steps are:

1. Access station selection;
2. MRS operational parameters configuration;
3. Tunnel connection re-establishment; and
4. MS CID mapping.

Steps 1–3 can each be skipped in a handover optimization while the last step, which is required for the RS to map the CIDs of the MSs it oversees, is not required for tunnel based forwarding.

An MRS handover can be initiated either by the MRS (through a MOB_MSHO-REQ) or the serving MR-BS (through a MOB_BSHO-REQ). If the MRS is switching MR-cells, the serving MR-BS may send the MRS's MAC address, in addition to the context of the MSs attached, to the target MR-BS. A target MR-BS may assign new CIDs or tunnel CIDs to the attached MSs; in this case, the MR-BS would inform the MRS through a RNG-RSP message of the old and new CID pairs so that the MRS would update its records and its forwarding.

The procedure described above does not involve a preamble change. The target MR-BS may decide that the MRS's preamble will change after handover. In this case, the target MR-BS sends the serving MR-BS a preamble index that is

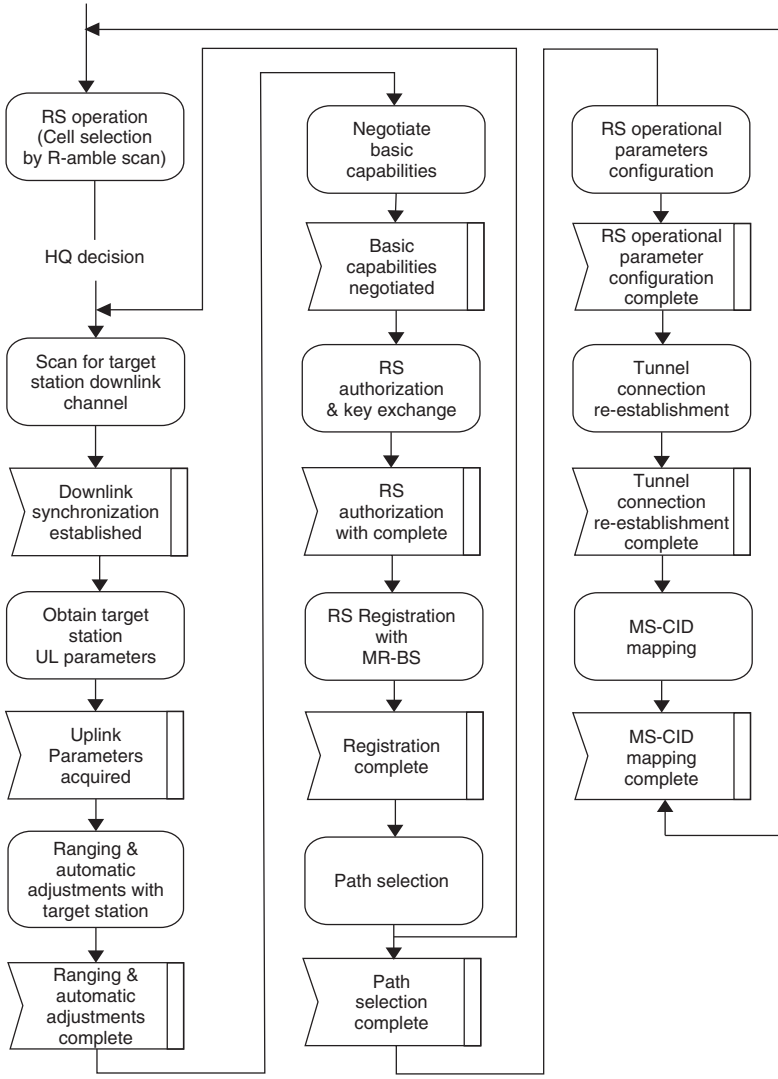


Figure 7.2 Flowchart for the mobile relay station handover process. Reproduced by permission of © 2009 IEEE.

forwarded to the MRS, including a frequency adjustment if required. The MRS would inform MSs attached to it of changes in channel characteristics through a MOB_NBR-ADV message that includes itself. Prior to the MRS handover, the serving MR-BS would have exchanged handover decisions and initiations with the MSs attached to the MRS.

For mobility within an MR-cell, an MR-BS may decide to change the preamble index of an MRS due to collisions in preamble index or to mitigate interference.

In such instances, the MRS would undergo a preamble index change process, and the attached MSs would be handed off from the MRS to itself using regular MS handover procedures.

If an MRS detects its drop during a handover, it will try to reconnect to its serving BS through cancelling the handover if still possible. If not, it attempts to reconnect with its preferred target MR-BS through reselection. If reselection fails, the MRS performs initial network entry procedure. In doing so, the MRS uses a HO code in its CDMA ranging. This enables the target MR-BS to recognize that the MRS's handover was not successful, and it may request the MRS's context (the MRS's and the attached MSs') from the serving MR-BS.

7.3 Mobility Management in IEEE 802.16m

The IEEE 802.16m amendment distinguishes between four types of handover:

1. Serving R1 BS to target R1 BS;
2. Serving ABS to target R1 BS;
3. Serving R1 BS to target ABS; and
4. Serving ABS to target ABS.

The first type is performed as per the IEEE 802.16-2009 standard. The amendment details description for types 2–4, in addition to inter-RAT handovers.

7.3.1 ABS to ABS Handovers

Similar to how legacy BS operate in IEEE 802.16-2009, an AMS acquires network topology either through periodic advertisements from the ABS or through scanning. An ABS advertisement contains information for neighboring ABSs and R1 BSs, but not neighboring CSG femtocells. A serving ABS may also unicast neighbor advertisement messages. In .16m, an AMS need not be assigned specific allocations by the serving ABS to perform scanning, and need not interrupt its communications with the ABS if such capability is supported. An AMS can prioritize the neighboring BSs to be scanned based on various metrics. Upon reporting these measurements to the network, either the AMS or the network can select a target BS to handover with. Conditions and rules for sending the AMS report are set by the ABS.

A handover can either be initiated by an AMS or commanded by an ABS, and either initiation or the command can include more than one Target-ABS (target ABS). If the ABS's command message contains only one target ABS, the AMS must adhere to this selection. An AMS's handover indication to the ABS results in stopping the serving ABS's downlink data and cancelling uplink allocations. If the command messages include more than one target ABS, the AMS would indicate its selection to the ABS. The serving ABS would define conditions with which the AMS would consider the target ABS(s) unreachable.

If all recommended ABSs are unreachable, the AMS would select a new ABS and indicate this choice to serving ABS.

There are three distinct phases to a handover procedure in the IEEE 802.16m amendment: initiation, preparation and execution. The amendment also defines procedures for handover cancellation.

Either the AMS or the ABS can initiate a handover. Handover conditions and triggers are defined by the serving ABS. An AMS's handover request begins the handover initiation, while an ABS's handover command begins both the initiation and the preparation phases.

The handover preparation phase is completed by the selection of a single target ABS. For example, a serving ABS command with a single target ABS completes the preparation phase. If the serving ABS's command includes more than one target ABS, the AMS's indication of the target ABS to the serving ABS completes the preparation phase. Preparation involves communication between the serving ABS and target ABS through the backbone to transfer context and to optimize the handover. A handover command signaling indicates which context information is transferred to the target ABS, in addition to information on the disconnect time with the serving ABS and the multiplexing schemes utilized if the AMS is to maintain simultaneous connections with the serving ABS and the target ABS during the handover procedure.

Handover execution starts at the time specified in the serving ABS command message. At that time, the AMS begins network re-entry procedures at the target ABS. If simultaneous communication is not supported, the serving ABS will stop downlink allocations at the disconnect time. Otherwise, the AMS stops communicating with the serving ABS once the network re-entry completes.

A handover can be cancelled at any phase during handover procedures. Cancellation would return both the serving ABS and the AMS to normal operations. Conditions for handover cancellation can be advertised by the network.

Network re-entry in IEEE 802.16m follows that of IEEE 802.16-2009. If a dedicated ranging code and/or a dedicated ranging channel are provided by the target ABS in the handover preparation phase, the AMS should utilize such setup during re-entry. CDMA-based handover ranging can be omitted when the AMS performs the handover to the target ABS.

7.3.2 Mixed Handover Types

Mixed handovers refers to when a handover takes place between a serving R1 BS and a target ABS, or from a serving ABS to a target R1 BS. In mixed these settings, network topology is acquired as follows. An R1 BS advertises the system information to its neighboring R1 BSs (per the IEEE 802.16-2009) and the LZones of its neighboring ABSs. An ABS advertises the system information for its neighboring R1 BSs in boths its MZones and LZones, the LZones system information for its neighboring ABSs in its LZone, and system information for its neighboring ABSs in its MZones. The ABS may indicate its Advanced capability through its LZone.

For a serving R1 BS to target ABS handover, the IEEE 802.16m amendment indicates that it will be possible for an R1 MS to handover to a target ABS's LZone using IEEE 802.16-2009 signaling and procedures. An AMS may also seek the same handover procedure as an R1 MS and switch zones (LZone to MZone) after handover. If the AMS is able to directly scan the Advanced-only target ABS's or the target ABS's MZone, it can perform such handover as well. At the moment, the IEEE 802.16m amendment does not detail how these handovers will be realized.

For a serving ABS to target R1 BS handover, R1 MS will proceed as if undergoing a IEEE 802.16-2009 handover. An AMS, however, would follow signaling and procedures of the Advanced system, but would perform network re-entry as per the IEEE 802.16-2009 procedures. A serving ABS would oversee the necessary mappings required between the Advanced and the IEEE 802.16-2009 for the context transfer.

7.3.3 Inter-RAT Handovers

An IEEE 802.16m network advertises information about other RATs through either solicitation or broadcast. The network acquires such information through a certain information server. Boundary information can also be broadcast by the IEEE 802.16m system through network boundary indication. An AMS, upon receiving this information, can perform measurements on the respective interfaces.

The IEEE 802.16m amendment discusses the possibility of generic handovers to variety other technologies such as 802.11, 3GPP and 3GPP2, but presumes that signaling details will be covered elsewhere, for example, the IEEE standard for Media Independent Handovers (802.21). The IEEE 802.16m amendment also discusses the possibility of enhanced inter-RAT handover procedures whereby the use of single or dual interfaces can be utilized.

7.3.4 Handovers in Relay, Femtocells and Multicarrier IEEE 802.16m Networks

For IEEE 802.16m with relay support, the ABS shall oversee the AMS handover procedures including scanning network topology advertisement. An ARS would only relay MAC control signaling between the AMS and the ABS. If the same AMS context is utilized between the ABS and its subordinate, no context transfer is necessary.

For systems with Femtocells, macrocell-Femtocell as well as Femtocell-Femtocell handovers are supported. A Femtocell going out of service due to network management instructions or by accident must initiate handover for its subordinate MSs to other macro or Femtocells. An MS should be able to prioritize the choice of accessible macro and Femtocells.

Handovers from macrocells to Femtocells would not be allowed to CSG-Femtocells if the MS is not a member, and to OSG unless it is critical for the MS's operation. In both cases, handovers are allowed in instances of emergency. Femtocell information can be advertised by the network and may be cached by the MS for future handovers. Information about CSG Femtocells are not broadcast, but either unicast or multicast to its members during handover preparation to a target Femtocell. Triggers and conditions for macro to Femtocell handovers are dictated by the network.

Meanwhile, for Femtocell to macrocell or other Femtocell BSs, network topology information can be either unicast or multicast depending on target BS accessibility to the AMS. Such handovers would proceed as a regular handover described above. Upon a successful Femtocell to macrocell handover, either the MS or network can cache handover information in case required for a reverse direction handover.

Under multicarrier operation, handover procedures are as described in Section 7.3.1. An ABS may broadcast/multicast/unicast its neighbors' multicarrier information to its subordinate AMSs. Management messages, however, are exchanged over the AMS's primary carrier. Network re-entry with the target ABS is performed on an assigned fully configured carrier at action time while fully communicating with the serving ABS. All primary and secondary carrier communication with the serving ABS cease once an AMS's network re-entry completes at the target ABS. An AMS capable of processing multiple carriers at the same time can seek a different primary carrier than the one specified in the serving ABS handover command. Once a handover is complete, network re-entry is performed through a target primary carrier. Once re-entry completes, the AMS may proceed to communicate over its primary and/or secondary carriers. Regardless of an AMSs multicarrier support, it may perform scanning and HO signaling with neighboring ABSs over multiple radio carriers while maintain normal operation with the serving ABS if the AMS is capable of concurrently processing multiple radio carriers.

8

Security

The IEEE 802.16-2009 standard describes a security sublayer that oversees entity authentication and message privacy, and defines the primitives required for these operations. The IEEE 802.16m amendment refines the IEEE 802.16 in security with additional security primitives and operations.

This chapter is organized as follows. Section 8.1 describes the Security Sublayer in IEEE 802.16-2009, including security associations, authentication mechanisms and encryption. In defining the IEEE 802.16j for multihop relay in WiMAX networks, the standard introduced the notion of Security Zones in order facilitate security management. This notion is described in Section 8.2. Finally, an overview of security in for the IEEE 802.16m is provided in Section 8.3.

8.1 Security in IEEE 802.16-2009

The IEEE 802.16-2009 security sublayer is shown in Figure 8.1. In essence, the security sublayer provides for two functionalities: encapsulation and key management. Encapsulation is achieved through a set of defined cryptographic suites that match data encryption techniques to authentication algorithms. Key management refers to how encryption and authentication keys are exchanged and updated during a connection's lifetime.

The standard describes the components of the security sublayer as follows:

- *PKM Control Management*: Controls all security components.
- *Traffic Data Encryption/Authentication Processing*: Encrypts/Decrypts traffic and relevant authentication functions.
- *Control Message Processing*: Process various PKM-related MAC messages.
- *Message Authentication Process*: Executes message authentication function.
- *RSA-based Authentication*: Performs RSA-based authentication function using the SS's X.509 digital certification and the BS's X.509 digital certification.

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

This stack is only engaged when RSA is selected as the authorization policy between an SS and a BS.
- *EAP Encapsulation/Decapsulation*: Provides interface with the EAP layer, when EAP-based authorization or the authentication EAP-based authorization is selected as an authorization policy between an SS and a BS.
- *Authorization/SA Control*: This stack controls the authorization state machine and the traffic encryption key state machine.
- *EAP and EAP Method Protocol*: Dependant on the usage of the upper layers, and is beyond standard's scope.

8.1.1 Security Associations

A Security Association (SA) is the basic security connection in IEEE 802.16-2009, and comprises a set of information that is shared between a BS and one or more of its client SSs. In a diversity handoff (MDHO or FBSS), this information can also be shared between the SS and BSs in the diversity set. During initialization, an SS established a Primary SA. Static SAs are maintained within the BS and Dynamic SAs are created on demand for the initiation and termination of service traffic flows. Both Static and Dynamic SAs can be shared by multiple SSs, for example, secure multicast. The contents of a SA include the SA's ID (SAID) that is unique to the SS, and key information required for traffic and signaling exchange in addition to their lifetimes.

Connections are mapped to SAs as follows.

- All transport connections shall be mapped to an existing SA.
- Multicast transport connections may be mapped to any Static or Dynamic SA.
- The secondary management connection shall be mapped to the Primary SA (however, no explicit mapping is required).
- Basic and primary management connections shall not be mapped to an SA.

In effect, and as the standard stipulates, "[a]ll MAC management messages shall be sent in the clear to facilitate registration, ranging and normal operation of the MAC." Note that an SS's Primary SAID equals the SS's Basic CID.

8.1.2 Authentication

There are two manners in which an SS can be authenticated by the BS. The first, called RSA authentication, involves the SS sending an X.509 certificate in its Authentication Information to the BS. The certificate, provide by SS's manufacturer, contains the SS's MAC address in addition to its public key. The second authentication type is EAP based and can utilize either the X.509 certificate or a Subscriber Identity Module (SIM) card provided by the operator. Support for RSA authentication is mandatory in the version 1 of the standard's PKM and optional in PKMv1. EAP authentication is only supported in PKMv2.

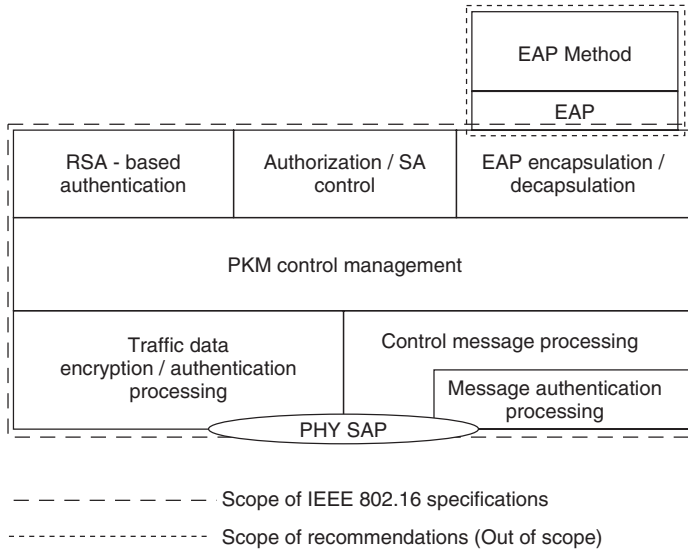


Figure 8.1 Security Sublayer. Reproduced by permission of © 2009 IEEE.

The intent of the Authentication Information is informational, and it may be discarded by the BS. An SS's Authorization Request immediately follows an Authentication Information. In addition to reiterating the SS's X.509 certification, the request contains a list of cryptographic suites that the SS support. In its Authorization Reply, a BS validates the SS's identity, determines cryptographic suite for operation, and provides Authentication Key (AK) for the SS. The authentication key is encrypted with the SS's public key. Both the Request and Reply include randomly generated number to ensure key liveliness.

PKMv2 allows for mutual authentication, which is not supported in PKMv1. Mutual authentication enables SSs to authentication their BSs. This RSA exchange involves the BS additionally providing its X.509 certificate for the SS in the Authorization Reply.

For handovers, the standard allows for a preauthentication process whereby an MS and a target BS would establish an AK prior to handover to accelerate reentry. The exact mechanism for preauthentication is beyond the standard's scope.

8.1.3 Encryption

Traffic Encryption Key (TEK) is acquired once authorization is achieved. A separate TEK is maintained by the SS for each SAID through making a Key Request from the BS. A BS would respond with Key Reply where the key would be encrypted by a Key Encrypting Key (KEK) that is derivable from the AK. The manner in which all keys are derived is defined in the standard, in addition to the relevant key hierarchies (i.e., RSA, RSA-EAP, EAP without RSA,

Table 8.1 Cryptographic suites defined in the standard. Reproduced by permission of © 2009 IEEE

#	Encryption	Data Authentication	Key Encryption
1	None	None	None
2	CBC mode 56-bit DES	None	3-DES,128
3	None	None	RSA, 1024
4	CBC mode 56-bit DES	None	RSA, 1024
5	CCM mode AES	None	AES, 128
6	CCM mode 128-bit AES	CCM mode, 128-bit	ECB mode AES with 128-bit key
7	CCM mode 128bits AES,	CCM mode	AES key wrap with 128-bit key
8	CBC mode 128-bit AES	None	ECB mode AES with 128-bit key
9	MBS CTR mode 128 bits AES	None	AES ECB mode with 128-bit key
10	MBS CTR mode 128 bits AES	None	AES key wrap with 128-bit key

CMAC/HMA/C from AK). The standard also describes the context for each key, how it is obtained and the scope of its usage.

At all times, the BS maintains two active sets of keying material per SAID where the lifetimes of the two keys overlap in order to maintain continuous encryption. A TEK is maintained as long as the SS's authorization is validated by the BS for the network, that is, active AK, and the SA.

Table 8.1 shows the cryptography suites as defined in the standard.

8.2 Security in IEEE 802.16j-2009

The IEEE 802.16j-2009 amendment describes two security control modes: centralized and distributed. Under centralized security control, an intermediate RS plays no role in any security exchange between the MR-BS and the SS. It is also possible that an SA be established between an MR-BS and RS. Similarly in such cases, intermediate RS do not intervene. However, it is optionally possible to protect non-authenticated PKM messages, such as Authorization Requests and Replies, through utilizing an added HMAC/CMAC between the MR-BS and the access RS.

Under distributed security control, two primary SAs are setup: one between the MR-BS and the access RS, the other between the access RS and the SS or the subordinate SS. In other words, an exclusive SA shall be established between each SS and its serving RS, and between each RS and its serving MR-BS, with each SA having its own SAID. An SS's management message is protected through replacing HMAC/CMAC values at the header. Note that it is possible for an RS to aggregate/deaggregate security messages for its subordinates in a single management tunnel.

The transfer of AK from the MR-BS to an SS or an RS would be made through a PKMv2 AK transfer message that also includes the AK's key material, sequence

number and lifetime. The amendment also allows for key pre-distribution to accelerate handoffs.

8.2.1 Security Zones

The amendment defines the management of security zones in MMR networks. In essence, a security zone consists of an MR-BS and a number of RS sharing a security context for the protection of relay management traffic. An RS becomes eligible to join a security zone by successfully being authenticated into a network and being provided the security zone key material by the zone's MR-BS. An RS cannot operate in a security zone before it joins or after it leaves. Security zone SAs, context, key usage and key derivations are all described in the amendment.

8.3 Security in IEEE 802.16m

Several changes and enhancements are introduced in the amendment for IEEE 802.16m. Figure 8.2 shows the counterpart for the security sublayer in .16m, called the Security Architecture.

Within the AMS and the ABS security architecture, there are two logical entites: the security management entity and the encryption and integrity entity.

The security management entity functions include:

- Overall security management and control;
- EAP encapsulation/decapsulation for authentication;
- Privacy Key Management (PKM) control (e.g., key generation/derviation/distribution, key state management);
- Authentication and Security Association (SA) control; and
- Location Privacy.

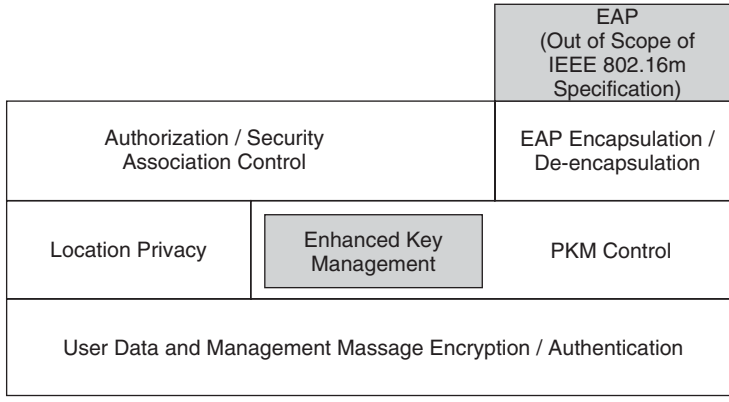


Figure 8.2 Security Architecture in 16m. Reproduced by permission of © 2009 IEEE.

The encryption and integrity protection entity functions include:

- Transport data encryption/authentication processing;
- Management message authentication processing; and
- Management message confidentiality protection.

However, certain noteworthy changes and enhancements are described below.

a. All authentications between AMS and ABS take place in EAP;
b. Only unicast static SAs are supported;
c. Station IDs are introduced whereby a station's MAC can be made private;
d. The introduction of Null SAIDs to accommodate mixed (secure/nonsecure) flow mixes;
e. Mapping multiplexed payloads onto SAs;
f. The possibility of encrypting MAC control message in three levels: no encryption, encrypted payload; encrypted payload and header.

Supported security modes for relaying include centralized and distributed.

For multicarrier communication, all security messaging is performed for the AMS's primary multicarrier.

Note that the draft does not detail how security is managed between Advanced and legacy systems.

Part Two

LTE and LTE-Advanced Networks

9

Overview of LTE and LTE-Advanced Networks

3GPP's Long Term Evolution (LTE) is a mobile broadband access technology founded as a response to the need for the improvement of to support the increasing demand for high data rates. The standard for LTE is a milestone in the development of 3GPP technologies. It came as an answer to the competition in performance and cost of IEEE 802.16-2009 to maintain the 3GPP systems share of the cellular communications market.

The chapter is organized as follows. Section 9.1 provides an overview of LTE and its successor, LTE-Advanced. It describes the protocol architecture, the connection interfaces, the support for femtocells and the air interface. Section 9.2 provides an overview of Part II of the book, going over frame structure, user equipment states and state transitions, quality of service management, mobility management and, finally, security.

9.1 Overview of LTE Networks

LTE is the Radio Access Network (RAN) of the Evolved Packet System (EPS). The network core component of EPS, called Evolved Packet Core (EPC) or System Architecture Evolution (SAE), is designed to be a completely IP-centric network that provides QoS support and ensures revenue and security. Figure 9.1 shows the basic architecture components of LTE, which consists of enhanced nodeBs (eNBs) at the RAN, and Mobility Management Entities (MMEs) and Serving Gateways (S-GW) at the core. The eNBs interconnect through an interface called the X2 interface, while they are connected to entities at the core (MMEs and S-GWs) using the S1 interface [1].

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

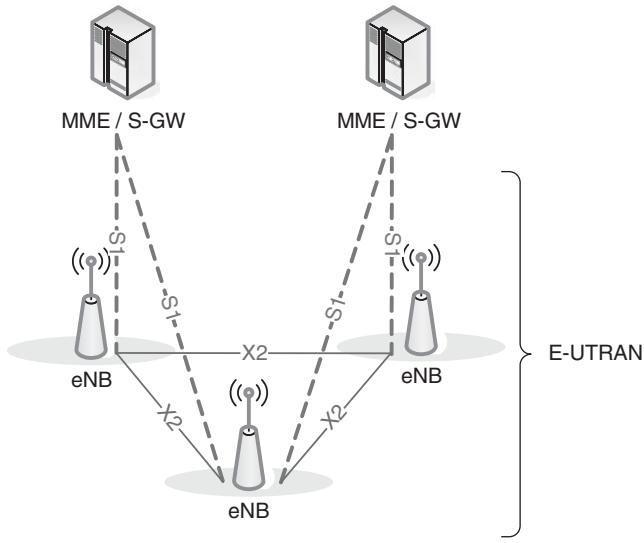


Figure 9.1 Basic LTE and LTE-Advanced Architecture. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

The LTE architecture depends on a network configuration that is simpler that its predecessor, the UMTS Terrestrial Access Network (UTRAN). In LTE, which is also called evolved UTRAN (EUTRAN), RAN considerations and decisions are all handled by the eNB, while relevant considerations for the core network are processed at the core. This "functional split", elaborated upon in Figure 9.2, directly results in substantial performance enhancements in cellular networks. The split further identifies the boundaries between the two network management and control stratums, where the Access Stratum (AS) is mostly handled by the eNBs and the Non-Access Stratum (NAS) is handled by the various entities at the core. Accordingly, an eNB would handle functionalities such as radio access control, scheduling, measurements at the radio interface, admission control, mobility control and inter-cell radio resource management. Entities at the core, including by the Mobility Management Entity (MME); the Serving-Gateway (S-GW); and the Packet Data Network Gateway (P-GW), would oversee functionalities such as mobility anchoring, NAS security, mobility while the User Equipment (UE) is in the idle state, and IP address allocation and packet filtering. It is also the network core that interfaces with other RANs and the Internet.

Figure 9.2 also shows the protocol stack at the eNB, including the PHY, MAC, Radio Link Control (RLC), Packet Data Convergence Protocol (PDCP) and the Radio Resource Control (RRC). The MAC, RLC and PDCP comprise the layer 2 protocols, while the RRC sublayer is a layer 3 protocol and is part of the control plane.

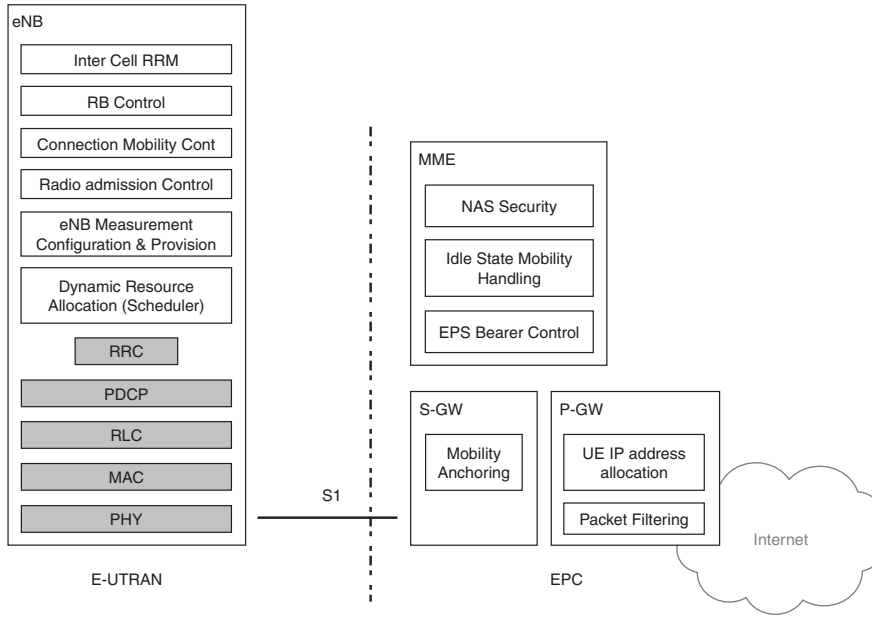


Figure 9.2 Functional split in LTE and LTE-Advanced. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

9.1.1 The Radio Protocol Architecture

LTE also maintains a 3GPP split between two protocol planes, the user plane and the control plane. The protocol stacks for the user plane is shown in Figure 9.3 while the protocol stack for the control plane is shown Figure 9.4.

The MAC performs several functions, including the mapping between logical and transport channels, multiplexing MAC SDUs, relaying scheduling information, error correction (HARQ), and priority handling. The RLC performs error correction (ARQ); concatenation, segmentation and reassembly for RLC SDUs; in addition to reordering and duplication detection. The PDCP, the higher sublayer in layer 2, mainly oversees ciphering and integrity protection, and transfer of control plane data.

The RRC is the RAN component of the control plane, and is responsible for the main control functionalities including broadcast of system information related to both the AS and the NAS, paging, establishing RRC connectivity between UE and the EUTRAN, security functionalities, mobility management functionalities, QoS management and transfer of NAS messages.

The NAS comprises all communication and signaling between the UE and the EPC that are relayed by the eNBs. The UE corresponds to the core only through the MME. The NAS performs many tasks including EPS bearer management,

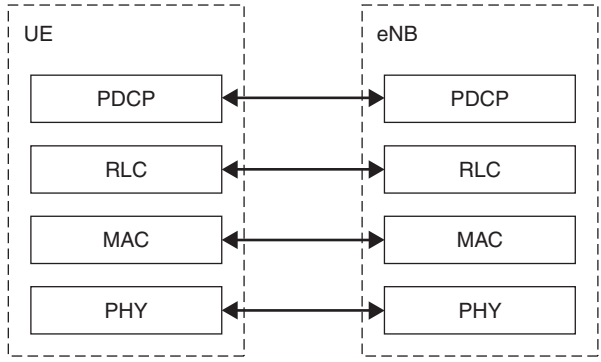


Figure 9.3 LTE User Plane. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

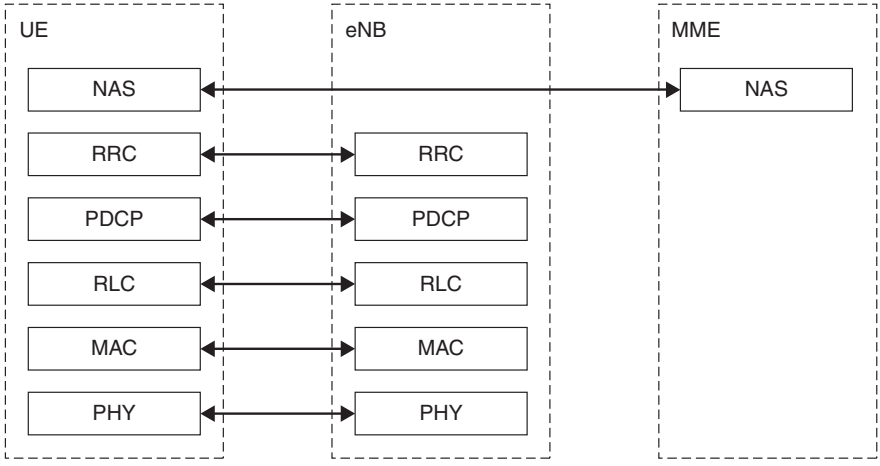


Figure 9.4 LTE Control Plane. Adaptation. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

authentication, paging and mobility management when the UE is in idle state, and security control.

9.1.2 The Interfaces

The S1 user plane, defined between the S-GW and the eNB, relies on the GPRS Tunneling Protocol (GTP) which, in turn, relies on User Datagram Protocol (UDP). The S1 control plane, on the other hand, is defined between the eNB and the MME, and utilizes the more reliable Stream Control Transmission Protocol (SCTP) for transferring signals. Through the S1 interface, the EPC performs the

main network management functions including radio access bearer management functions; mobility functions in instances of intra-LTE and inter-3GPP RAT handovers; paging; user context setup, management and transfer, and load balancing between MMEs.

The X2 interface, used to interconnect eNBs, also has user and control planes. As in the S1 control plane, the user plane in X2 delivers user plane data in a non-guaranteed fashion based on a GTP/UDP stack, while control plane signaling depends on SCTP for reliable delivery. The control plane functionalities overseen by the X2 interface include intra-LTE mobility support (including context transfer and control of user plane tunnels between serving eNB and target eNB), load management between eNBs, and general X2 management and error handling functions.

9.1.3 Support for Home eNBs (Femtocells)

3GPP Release for LTE showed a clear support for Home eNBs (HeNBs) or femtocells. A femtocell connects to the EPC through the S1-MME and S1-U interfaces. It is possible that a HeNB gateway be employed to allow the S1 interface between the HeNBs and the EPC to scale and support a large number of HeNBs. The HeNB gateway (HeNB GW) would appear to a HeNB as an MME, while for the MME the gateway would appear as a HeNB. Whether a HeNB connects to the EPC directly or not, the S1 interface remains the same. An EUTRAN with HeNB is shown in Figure 9.5.

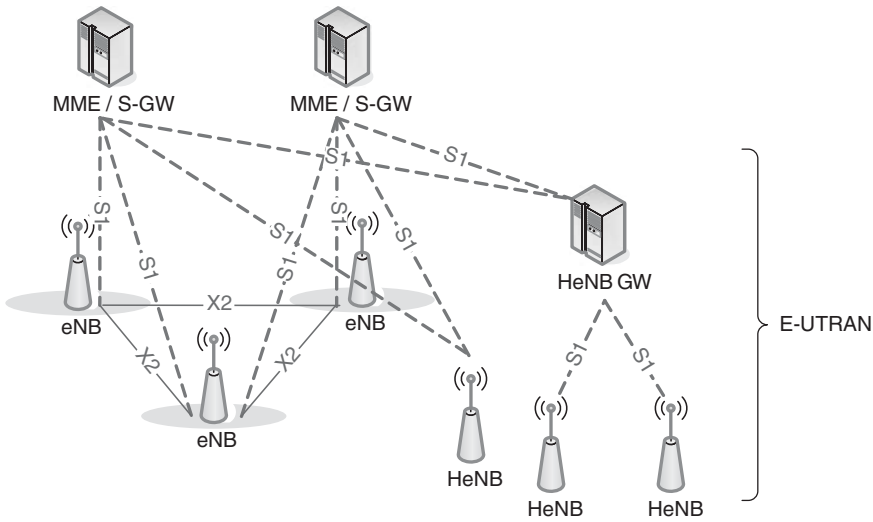


Figure 9.5 Architecture with HeNBs. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

HeNB has the same protocol stack as a regular eNB, but performs additional functionalities such as Serving HeNB gateway discovery, in addition to access control. In turn, the HeNB gateway relays S1 signaling to the core. Non-UE signaling also terminates at the HeNB gateway. It should be noted, however, that an X2 interface is not defined for HeNB in as of Release 9.

The above noted access control depends on whether an HeNB is closed, open or hybrid. If it is closed, that is, CSG, then only users in the CSG can access and utilize the resources of this HeNB. On the contrary, an open HeNB, that is, OSG, is accessible to all EUTRAN users. Finally, a hybrid HeNB, as the name suggests, is one with a CSG, yet allows open access if sufficient resources are available (i.e., non-subscribers can be admitted if their requirements do not adversely affect those of the subscriber group). Note that in all instances, it is possible to admit emergency calls into a HeNB.

9.1.4 Air Interface

LTE and LTE-Advanced use OFDM as the PHY modulation method, and employ OFDMA as the multiple access scheme for the downlink communication [2]. LTE and LTE-Advanced supports both duplexing modes TDD and FDD. LTE was the first technology to support both TDD and FDD, while WiMAX initially supported TDD, and then extended its support to FDD in later amendments. Employing OFDMA to access the downlink air interface based on allocating physical resource elements in frequency and time dimensions called Physical Resource Blocks (PRBs). Allocating PRB incur complexity over simpler access methods, like plain TDD or FDD, in terms of more involved scheduling. The scheduler resides at the BS side. At each frame, the scheduler allocates specific number of PRBs to each active user in an attempt to meet the traffic demand of all the active users.

The frame duration in LTE and LTE-Advanced networks is 10 ms. For FDD, the whole frame is used for downlink transmission; however, for TDD the frame is divided into uplink and downlink transmissions. Each frame consists of 10 subframes with 1 ms duration each. The subframe is further divided into two time slots of 0.5 ms duration each. Furthermore, the standard specifies two CP lengths, a short CP with 4.7 μs (used for small cell coverage) and a long CP with 16.7 μs (used for large cell coverage). Consequently, if the short CP is used, a time slot would consist of seven OFDM symbols, while if the long one is used, it would consist of six symbols. By definition, an PRB is a single time slot spanning 12 frequency subcarriers. As a result, the number of available PRB in an LTE system depends on the downlink channel bandwidth. The standard specifies a number of possible channel bandwidths, namely 1.25, 2.5, 5, 10, 15, and 20 MHz, where the corresponding numbers of PRBs are 6, 12, 25, 50, 75, 100 respectively. The amount of data rate (in bits per second) available at the downlink channel between a BS and an UE depends on the type of modulation

employed. LTE supports different type of modulation such as PSK, QPSK, and 64 QAM.

LTE-Advanced relies on the architecture of LTE/SAE, but utilizes additional advancements at various levels. In terms of network structure, LTE-Advanced will utilize the relaying functionality whereby the relay station would connect to the core through a donor cell, that is, a Donor eNB (DeNB).

There are several ways to classify relaying in LTE-Advanced. The first is based on spectrum usage. Relaying can be either in-band or out-of-band, where in the former the eNB-relay link shares the same carrier frequency with the relay-UE links while in the latter the two sets of links do not operate in the same carrier frequency. Relaying can also be classified based on the UE's awareness of the relay station's existence, that is, whether the relay station is transparent or non-transparent. The transparency dictates whether the relay is part of the donor cell, or controls a cell on its own. When smart repeaters, decode-and-forward, L2 relays and Type 2 relays are used, the relay does not have an identity of its own. Some RRM functionalities, however, may reside at the relay station. Meanwhile, when L3 or Type 1 relays are used, the relay station is uniquely identified by the UE. The relay station may also carry the full suite of RRM functionalities as an eNB.

Classifying relays into Type 1 and Type 2 is similar to that of non-transparent and transparent, respectively. Type 1 relays are in-band relays uniquely identified by the UE, communicate directly and fully with the UE (uplink and downlink), and appear as a Release 8 eNodeB for Release 8 EUs. The further classification of Type 1 relays into Type 1a and 1b distinguishes isolation type. Type 1a relays are out-of-band relays while Type 1b are in-band relays with antenna isolation. Type 2 relays are not uniquely identifiable by the UE and are transparent to Release 8 UE. For current considerations, at least Types 1 and 1a are to be supported in LTE-Advanced.

9.2 Overview of Part II

9.2.1 Frame Structure

Chapter 10 will discuss the frame structure in both LTE and LTE-Advanced. In TDD, only one carrier frequency is used for uplink and downlink transmission where the uplink and downlink transmissions are only isolated in time. The TDD frame, also known as type 2 frame, is divided between the two transmissions using a guard period which is required to switch between the them. In every frame, either one or two sub-frames are split into a downlink part called DwPTS, a guard period, and an uplink part called UpPTS to facilitate the switching from downlink to uplink or vice versa. Other sub-frames are either allocated to uplink or downlink transmissions. The only exception is that sub-frames 0 and 5 are always allocated to downlink transmission.

9.2.2 UE States and State Transitions

3GPP extensively defines procedures for camping and network entry in LTE and LTE-Advanced. These are described in Chapter 11.

In LTE, a UE is required to perform certain steps in order to join a specific cell after being switched on. These steps compose the initial access procedure according to the LTE terminology. In particular, these are:

1. Cell search and cell selection;
2. Receiving system information; and
3. Random access procedure.

Cell search is the process by which a UE acquires time and frequency synchronizations with a cell and detects eNB's ID. Acquiring the cell ID is by itself a two-stage process. First, the cell ID needs to be identified, and second the cell ID group is acquired. The standard identifies 504 PHY cell identities, which are divided hierarchically into two tiers: a 168 unique cell layer identity groups, with each group comprising three physical layer identities. The BS broadcasts the PHY identity over a signal called the Primary Synchronization Signal (PSS), and the cell layer identity group in the Secondary Synchronization Signal (SSS). Besides these signals, LTE BS transmits cell-specific reference signal that contains the downlink channel estimation for coherent demodulation and Channel Quality Indicator (CQI). The UE first looks for the PSS. This signal is normally transmitted in the last OFDM symbol of the first time slot of sub-frame 0 and sub-frame 5. The PSS enables the UE to get time synchronized on 5 ms scale. After receiving and analyzing the PSS signal, the UE obtains the radio frame timing and the cells' group identity from the SSS. The SSS signal is also transmitted in the 0 and 5 sub-frames. Hence, it is periodically received every 5 ms, which assists the UE to fully synchronize with the BS.

After successful cell search and selection, the next step is getting the system information. The UE configures the logical Broadcast Control Channel (BCCH) via the Broadcast Channel (BCH) and maps it to the Physical Broadcast Channel (PBCH) to be able to decode the Master Information Block (MIB). The MIB contains indispensable system information for the UE operation. MIB includes information about the number of available resource blocks, configuration of the Physical HARQ Indicator Channel (PHICH) and the System Frame Number (SFN). The MIB is transmitted in the first four OFDM symbols in the second time slot of sub-frame 0 and is repeated every fourth frame. After receiving the MIB, The UE reconfigure the BCCH channel and maps it to the PDSCH to receive System Information Block Type1 (SIB1), which includes the Physical Cell ID and scheduling information about other System Information Blocks Such as SIB2, SIB3, . . . etc. After receiving the SIB1, The UE uses the scheduling information about the SIBx ($x = 1, 2, 3 \ldots$) to extract information about the SIB2, mainly in which subframe SIB2 is transmitted and reconfigures the BCCH to receive the SIB2. The SIB2 includes the common channel information, random access

channel information and the random access preamble information, in addition to the information required for the HARQ procedure.

After receiving the system information, the UE is ready to start the random access procedure using the Random Access Channel (RACH) and common shared channel information to configure both channels. The UE starts the random access procedure to get its first allocated slots to transmit its uplink data for the first time. The UE contend on the uplink shared channel to send Random Access Preamble. If the preamble transmission is successful, the BS responds by the random access response. This response carries the Cell Radio Network Temporary Identifier (C-RNT) and uplink grant. After receiving the response, the UE can send connection requests to the BS on the CCCH to establish data connections to transmit its uplink data.

9.2.3 Quality of Service and Bandwidth Reservation

Chapter 12 reviews procedures for QoS management in LTE and LTE-Advanced. The LTE QoS framework is designed to provide an end-to-end QoS support. To achieve this, LTE provides QoS based on each flow requirements. LTE classifies flows into Guaranteed Bit Rate (GBR) GBR and non-GBR flows. Flows in LTE are mapped into radio bearers which are the over-the-air connections. To accommodate end-to-end QoS; LTE differentiates between two types of radio bearers, S1 bearers and EPS bearers. An S1 bearer is a connection between an eNB and either the MME or the S-GW, while an EPS bearer is a connection between the EPS and the MME or S-GW, or the S-GW and the P-GW. There are two types of bearers in LTE, default bearers and dedicated bearer. The former, which is a non-GBR bearer that does not provide bit rate guarantees, is initiated and established at the startup time to carry all traffic. On the other hand, the latter can be either a GBR or a non-GBR bearer. If it is a GBR bearer, the UE can specify the guaranteed bit rate, packet delay and packet loss error rate. Each dedicated bearer is characterized by a Traffic Flow Template (TFT) with QoS parameters associated to it. An uplink TFT is used to map the UE uplink Service Data Flow (SDF) to specific QoS parameters, with the mapping carried out at both the eNB and the UE. Mapping for the downlink TFT is carried out at the S-GW or the P-GW. LTE groups bearers into classes. Each class is identified by a scalar number called the QoS class Identifier (QCI). A QCI identifies a group of QoS parameters describing the packet forwarding treatment in terms of priority, tolerated delay, and packet error rate. Packet forwarding treatment is enforced by allocating radio resources for bearers through scheduling.

The scheduler resides in the eNB to dynamically allocate uplink and downlink resources over the uplink and downlink shard channel U-SCH and D-SCH, respectively. The uplink and downlink schedulers are invoked to allocate resources every Time Transmission Interval (TTI). The minimum TTI duration is of one subframe length; that is, 1 ms. uplink scheduling is performed per SC-FDMA PRB while downlink scheduling is performed per OFDMA PRB. eNB calculates the time-frequency resources given the traffic volume and the

QoS requirements of each radio bearer. However, the resources are allocated per UE and not per radio bearer.

In addition to the dynamic allocation, LTE provides the flexibility to what is called persistent scheduling where the time-frequency resources can be implicitly reused in the consecutive TTIs according to a specific periodicity. Persistent scheduling reduces the overhead scheduling for applications such as VoIP. Scheduler design is not specified in the standard and is left for vendor implementation. An efficient scheduler, however, should take into account link channel quality and the buffer length of the radio bearers. It should also cater to fairness among the UEs based on their Service Level Agreements (SLA).

The operation of HARQ is highly related to the scheduling operation. LTE provides two mechanisms of error detection and correction through re-transmission namely, the HARQ mechanism at the MAC layer and the ARQ at the RLC layer. The ARQ functions less frequently than the HARQ and handles errors not detected by the HARQ process. HARQ is designed to be simple and fast to improve the QoS performance. This improvement is achieved by reducing delay and increasing the system throughput through the fast retransmission. The feedback signal of HARQ is a one bit ACK/NACK and the HARQ can be sent at every TTI.

LTE-Advanced carrier aggregation has an impact on both scheduling and HARQ. For HARQ, it is required in carrier aggregation [3] whether contiguous or non-contiguous, to have one independent HARQ entity per scheduled component carrier. Note that the maximum number of HARQ entities allowed by LTE-Advanced is eight entities for the FDD duplexing. For scheduling, similar to Release 8, each UE may be simultaneously scheduled over multiple component carriers. However, at most one random access procedure is scheduled per UE in any time frame. For TDD, it is required that the number of component carriers of the uplink should be equal to that of the downlink. As in LTE, a single component carrier is still mapped into one transport block.

Relaying in LTE-Advanced defines two types of HARQ and two types of scheduling: end-to-end and hop-by-hop HARQ and centralized and distributed scheduling. The end-to-end HARQ is simple because the eNB has full information about the status of each HARQ transmitted block. HARQ is performed at the eNB and the UE, the RS only relays data and control message between the two. In hop-by-hop HARQ, the RS not only forwards the data from/to eNB/UE, but also contributes in processing. For example, when a RS receives a message from the eNB destined to the UE, the RS decodes the message, checks the Cyclic Redundancy Check (CRC) and generates its own feedback (ACK or NACK).

In Relay LTE-Advanced centralized scheduling, eNB is responsible for scheduling all links of the network, relay links and UE links over the one and two hops distance of the network. The RS only forwards the received data and signaling from eNB without any processing. In LTE-Advanced relaying networks employing distributed scheduling, the scheduler resides at both the eNB and the RS. The eNB only schedules resources for the eNB-RS links as well as for UEs connected directly to it. On the other hand, the RS schedules resources for

RS-UE links that are two hops away from the eNB. Consequently, the eNB does not need to receive the Channel State Information (CSI) of the RS-UE link, which makes distributed scheduling consume less signaling and overhead. Distributed scheduling can only be employed in Type I relaying networks.

9.2.4 Mobility Management

3GPP Mobility management is described in Chapter 13. LTE supports various users' mobility by standardizing handover essential signaling and processes [4]. There are three handover types in LTE:

1. *Intra-Handover*: Occurs within the same LTE network nodes (intra-MME and intra-S-GW).
2. *Inter-Handover*: Occurs between different LTE networks nodes (inter-MME and Inter-S-GW).
3. *Inter-RAT*: Occurs between different radio technology networks, for example WiMAX and LTE, UMTS and LTE, etc.

Also, there are two types of handover decisions based on the decision-making entity. The first is the network evaluated decision, where the eNB makes the decision while the second is the mobile evaluated handover, where the UE takes the decision of handover and conveys it to the serving eNB. In this type, the eNB can decide to either, meet or deny this request based on the current network conditions.

Intra-Handover

Intra-Handover is performed to handover a UE from a serving eNB to a target eNB over the X2 interface with the same MME and serving gateway. In general, LTE Intra handover processes include procedures to measure downlink channel quality between the eNB and the UE, procedure to process the channel quality data collected by UE, which is done by the UE, procedure to send the channel quality processed data from the UE to the serving eNB, and finally procedures used by the eNB to make a handover decision based on the data received from UE. The above mentioned procedures are grouped under one stage of the handover process called handover preparation which is the first of three intra-handover stages defined by LTE. The other two stages are handover execution and handover completion. The functionalities of these three stages are summarized next:

1. *Handover Preparation*: In this stage, the UE, the serving eNB and the target eNB are all involved with specific tasks performed by each. First, the UE processes and prepares a channel quality measurement report and sends it to the serving eNB. To carry out this task at the UE, the serving eNB should configure and trigger the UE measurement procedure. Once the serving eNB receives the measurement report, the serving eNB processes the report to come

up with a handover decision based on the level of the channel strength. If the serving eNB decides to hand the UE over to a target eNB, it sends a handover request to the target eNB. When the target eNB receives this request, it performs an admission control request based on the resources indicated to be required for the UE's radio bearers. If sufficient resources are available, the target eNB acknowledges the request with a handover request acknowledgement; otherwise it sends a handover preparation failure message to the serving eNB. This terminates that handover preparation procedure. If beyond a certain time the serving eNB does not receive an indication of either an acknowledgement or a failure, it sends a handover cancel and indicates the cause as expired timer. If a serving eNB sends a handover cancel requests, it disregards further messages from the target eNB. If the handover request acknowledgment is received, the serving eNB sends the handover command to the UE which includes all the necessary information for the UE to access the target cell. This finalizes the handover preparation stage.

2. *Handover Execution*: The UE uses the information included in the handover command to execute the handover process. UE performs different tasks during this stage. It performs the random access procedure over the RACH to connect to the target cell. Also, it acquires time synchronization with the target eNB. Timing advance for the UE is performed at the uplink and the handover confirmation message is given to the target eNB by the UE. This step is important to maintain the frequency subcarrier orthogonality necessary to mitigate intra-cell interference.
3. *Handover Completion*: This is the last stage, where the target eNB sends a confirmation and path switch request. A request and response for modifying the UE's radio bearers are then processed by the serving GW and, in turn, the PDN GW. Once completed, the serving GW responds to the path switch request by redirecting the UE's downlink data to the target side, sending an end marker to the serving eNB. In turn, the serving eNB sends another end marker to the target eNB and the MME acknowledges the target eNB's path switch request. Finally, the target eNB sends a context release message to the serving eNB. Upon reception of the context release message, the serving eNB releases control and data connection resources. This finalizes the handover procedure.

Note here that if the handover cannot be initiated over an X2 interface, the S1 mobility management oversees the handovers.

Inter-Handover

This type of handover is initiated over an S1 interface when the UE roams between two different MME areas; that is, the serving eNB and the target eNB are controlled by two MMEs with the same S-GW or two MMEs and two different S-GWs: serving MME-serving S-GW and target MME-target S-GW. Inter-handover is similar to intra-handover over S1 interface except for the involvement of two MMEs.

The handover is initiated by sending a handover required message by the serving eNB. This initiates an S1 handover preparation phase. The serving MME detects that the target eNB is in another MME. Hence, it forwards the request to the target MME. The target MME creates the S1 logical connection toward the target eNB and sends the S1 HANDOVER REQ over it. If the target MME judges that the handover can be realized, it sends a handover command. The target MME performs a handover resource allocation by sending a handover request message to the target eNB. Upon receiving the request message, the target eNB makes the appropriate resource allocation, context preparation and relevant security authentications for the UE.

downlink data packets are forwarded from the serving-eNB to target-eNB via the S-GW during the handover if the S-GW remains the same. An indication of a successful handover is sent by the target eNB to the target-MME using a handover notify message.

Inter-RAT Handover

Inter-RAT handovers will be discussed at length in Part-III of the book. However, similar to inter and intra handovers, an inter-RAT handover consists of three stages: preparation, execution and completion. However, the procedure in each stage differs by the type of technology involved in the handover with the LTE network. LTE and LTE-Advanced generally defines the signaling required to perform Inter-RAT handover, and leaves all other issues for the IEEE 802.21 standard, which defines mechanisms for Media Independent Handovers.

Handover in Femtocells

LTE standard defines three types of handovers in femtocells [5]:

1. *Inbound*: Handover from macrocell to femtocell. It is similar to handover from a macrocell to macrocell over S1 interface. In this handover, the serving eNB recognizes that the target cell is HeNB from the Tracking Area Code (TAC) and HeNB ID. Using this information, it identifies the HeNB gateway and sends the handover request to MME, which will forward the request to HeNB gateway. In turn, the HeNB gateway forwards the request to the target HeNB.
2. *Outbound*: Handover from femtocell to macrocell. The HeNB gateway receives the handover request from the serving HeNB and forward it to the MME over S1 interface. MME forwards the request to the target eNB. Inbound and Outbound handovers are expected not to be soft handovers due to the limitations of the frequency resources at the femtocells.
3. *Inter-Femtocell*: Handover between femtocells. The HeNB takes care of the femtocell to femtocell handover over the S1 interface.

Handover in Relay LTE

Handover procedures in relay LTE is changed by the introduction of the RSs. There are two types of handover processes in relay LTE: centralized and

distributed. The centralized process is almost similar to the handover of LTE except the introduction of the RS. In centralized process, the handover request is initiated by the serving eNB. The relay transparently forwards the measurement reports and requests from the UE. The serving eNB and the target eNB are in control of the handover process with the assistance of the RSs forwarding messages to the UE transparently.

In distributed process, the serving RS initiates the handover. The handover procedures are carried out with the collaboration of the RS (serving and target) to successfully conduct the handover.

Distributed process is similar to the LTE handover process with the difference that the serving and the target RSs are in control of the handover procedures besides the corresponding eNBs. In this type of handover, the serving RS receives the measurement report from the UE and forwards it to the serving eNB. Meanwhile, in centralized handover, this report is directly sent to the serving eNB. The admission control is carried out by the target eNB on the backhaul link and the target RS on the relay link in distributed handover, while it is performed by the target eNB on the backhaul and relay link without intervening of the target RS. Synchronization and timing advance for the UE is performed at the uplink and the handover confirmation message is given to the target RS by the UE. The target RS sends the confirmation message to the target eNB, which finalizes the handover procedure as per LTE handover procedures. Finally, the target eNB sends request to the serving eNB to release network resources. Finally, this latter forwards the command to the serving RS to release resources of the UE.

9.2.5 Security

Chapter 14 describes the security architectures and procedures in 3GPP. The separations of user and control planes and the access and NAS in LTE/SAE result in an implicit security requirement. LTE establishes security association with access stratums between the UE and eNB only if the UE is connected. However, if a UE is in idle mode, the eNB does not preserve states about a UE in idle mode. In UE idle mode the NAS messages are still exchanged. Hence, non-stratum security associations are established between the UE and the MME.

3GPP describes an extensive two layer security architecture that also utilizes the Internet Engineering Task Force (IETF) security solutions for its IP core. The security architecture is maintained in LTE-A, with some enhancements concerning more capable encryption and integrity algorithms being utilized.

There are five sets of security feature groups defined in LTE:

1. *Network access security*: The set of security features that provide users with secure access to services, which in particular protect against attacks on the (radio) access link.
2. *Network domain security*: The set of security features that enable nodes to securely exchange signaling data, user data (between the Access Network

(AN) and the Serving Network (SN), and within the (AN), and protect against attacks on the wireless network.
3. *User domain security*: The set of security features that secure access to UEs.
4. *Application domain security*: The set of security features that enable applications in the user and in the provider domain to securely exchange messages.
5. *Visibility and configurability of security*: The set of features that enables the user to inform himself whether a security feature is in operation or not and whether the use and provision of services should depend on the security feature.

EPS Authentication and Key Agreement (AKA)

The EPS AKA produces key hierarchy material forming a basis for the user plane, RRC and the NAS ciphering keys as well as RRC and NAS integrity protection keys. These keys are used to protect user plane traffic between the UE and the network. The key hierarchy is derived using cryptographic functions. It includes the following keys K_{eNB}, K_{NASint}, K_{NASenc}, K_{UPenc}, K_{RRCint} and K_{RRCenc}, which are respectively the keys for the eNB, the NAS traffic without encryption, NAS traffic with encryption, User Plane traffic with encryption, RRC traffic without encryption and RRC traffic with encryption. Keys for unencrypted traffic are used to verify integrity.

The main advantages of key hierarchy and cryptographic derivation are:

1. If one attacker gets hold of one key, he cannot be able to generate other keys because they are at an upper layer in hierarchy.
2. Keys are bound to the location and purpose they are used for. This prevents using a key in different access networks if this key is compromised; that is, key used in one AN cannot be used in another.
3. Keys used between UE and eNB are changed regularly (e.g. during the initial network entry or handover) without a need to change the root key.

The MME sends to the Universal Subscriber Identity Module (USIM) a random challenge, an authentication token, in addition to the K_{ASME}. The K_{ASME} key is a base key, from which NAS keys, K_{eNB} keys and H are derived. The K_{ASME} is never transported to an entity outside of the EPC, but K_{eNB} and NH are transported to the eNB from the EPC. From the K_{eNB}, the eNB and UE can derive the UP and RRC Keys.

Handover Security

When handover is initiated by the serving eNB, it is required to transfer security parameters to the target eNB in a trusted environment. To achieve this, LTE introduces the concept of forward security as the property that, for an eNB with knowledge of a K_{eNB}, shared with a UE; it shall be computationally infeasible to predict any future K_{eNB} that will be used between the same UE and another eNB. Hence, it is infeasible for an attacker who compromised either, a serving or

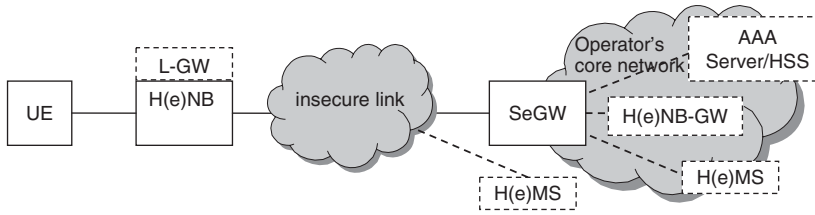


Figure 9.6 Security system architecture of HeNB.

a target eNBs and obtained its key, to deduce or know the key of the other one. Meanwhile, the UE has full information to deduce the required key. In case a target eNB is compromised during handover (backward security), LTE defines a procedure to ensure keeping previous traffic secured. In this case a serving eNB derives a new key from current key and only transfer this key to the target eNB.

LTE Relay Security

Current LTE standard do not specify or resolve security issues on the relay link and the backhaul link. However, it provides general guidelines for LTE relay security. LTE relay assumes the RS has secure environment for forwarding and processing data. It mandates mutual authentication between RS and network using AKA and RS device authentication. The standard requires binding between these two authentications procedures. It necessitates the control plane traffic to be integrity protected while the user plane traffic integrity protection is left optional. Other mandates of LTE relay security is the confidentiality protection between the RS and the network. Until March, 2011, work on LTE relay security is ongoing to evaluate solutions and procedures before standardization.

LTE Femtocell Security

Figure 9.6 shows the security architecture of HeNB [6]. As the figure shows, HeNB accesses the core network via Security Gateway (SeGW), while the link between HeNB and SeGW may be insecure. Once HeNB accesses the core network, SeGW performs a mutual authentication with the HeNB. If a HeNB is authenticated, a security tunnel is established between it and the SeGW to protect information transmitted in backhaul link. The security tunneling protocol can be IPsec or any other layer two security protocol even though LTE standard mandates implementation of the security channel. However it leaves using it optional based on an operator policy. In the CSG case, the HeNB-GW performs the mandatory access control, while HeNB performs the optional access control in case of OSG HeNBs.

Different security procedures are defined for HeNB security system, mainly:

1. Device integrity check performed by HeNB and the Trusted Environment (TrE) upon booting and before connecting to the core network and/or to the

HeNB Management System (H(e)MS). TrE is a logical entity which provides a trustworthy environment for the execution of sensitive functions and the storage of sensitive data.
2. Device validation and authentication of the HeNB device platform, where a device implicitly indicates its validity to the SeGW or HeMS by successful execution of device authentication.
3. Device Authentication may optionally be followed with an EAP-AKA-based hosting party authentication exchange.
4. As a result of the authentication procedure, IPsec tunnel establishment by HeNB. At least one IPsec tunnel is set up to protect all signaling, user, and management plane traffic over the interface between H(e)NB and SeGW.
5. Optionally an AAA server may be used to verify the authorization of the H(e)NB to connect to the operator's network based on the authenticated device identity extracted from the H(e)NB certificate.
6. Location verification performed by the H(e)MS and/or HNB-GW to ensure the H(e)NB location satisfies various security, regulatory and operational requirements.

References

[1] 3GPP TS 36.300 V10.2.0 (2010-12) (Release 10), Technical Report, Overall description; Stage 2.
[2] 3GPP TR 36.814 V9.0.0 (2010-03) (Release 9), Further advancements for E-UTRA PHY aspects.
[3] 3GPP TR 36.808 V1.0.0 (2010-12) (Release 10), Carrier Aggregation Base Station (BS) radio transmission and reception.
[4] 3GPP TS 36.133 V10.1.0 (2010-12) (Release 10), Requirements for support of radio resource management.
[5] 3GPP TS 25.367 V9.5.0 (2010-12) (Release 9), Mobility procedures for Home Node B (HNB); Overall description.
[6] 3GPP TS 33.320 (2010-12) (Release 11), System Architecture of H(e)NB.

10

Frame-Structure and Node Identification

3GPP Release 9 describes the frame structure for LTE, in addition to mechanisms and procedures for naming and identifying network entities. Substantial differences are observed in Release 10, which describes the LTE-Advanced networks. These include changes in the frame structures to support coexistence with legacy elements and other IMT-Advanced networks, support for relay networks, in addition to supporting femtocells.

This chapter discusses the frame structure and identification and addressing in both LTE and LTE-Advanced. It is organized as follows. In Section 10.1, the frame structure for LTE and the structure of the resource block is identified. In Section 10.2 the frame structure for LTE-Advanced is introduced. Section 10.3 is concerned with the identification, naming and addressing of various entities in 3GPP networks, based on Releases 9 and 10.

10.1 Frame-Structure in LTE

In LTE, DL and UL transmissions are organized into radio frames of 10 ms each. Each frame is divided into ten equally sized subframes. The duration of each subframe is 1 ms. Moreover, each subframe is further divided into two equally sized time slots, that is, each slot is 0.5 ms. 3GPP defines two types of frames based on the duplexing scheme used. These are Type 1 when FDD is used and Type 2 when TDD is used. Figures 10.1 and 10.2(a) illustrate the two types, respectively.

In Type 1 frames, DL and UL transmissions use two different frequency bands. Hence, frames are not shared between the two. On the other hand, in TDD, the two transmissions share the same frequency bands but are separated in time. Hence, they share the frames. In fact, every frame is divided into two halves, one for the DL transmission while the other is for the UL transmission. Nevertheless,

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

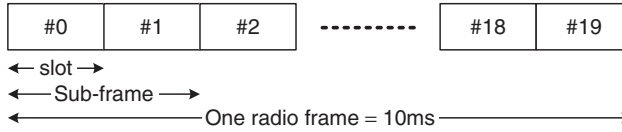


Figure 10.1 Type-1 Frame structure. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

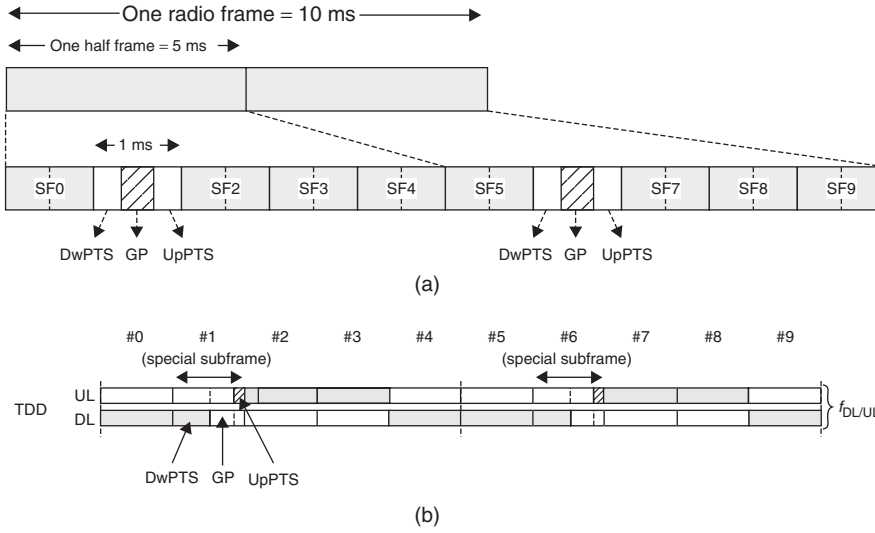


Figure 10.2 Type-2 Frame structure. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

a Type 2 frame is similar in structure to a Type 1 frame. The only difference is the existence of one or two special subframes that help switching between UL and DL transmissions.

A special subframe consists of three fields: a Downlink Pilot Time Slot (DwPTS), an Uplink pilot time slot (UpPTS), and a Guard period (GP) in between the two (see Figure 10.2(b)). The lengths of the DwPTS and UpPTS are configurable, but are constrained (together with the GP) by a total fixed length of 1 ms, that is, the duration of one subframe. The DwPTS can be considered as an ordinary DL subframe, that is, 1 ms, and can be used for DL transmission. It may also be of a shorter duration, as it can vary from three to twelve OFDM symbols. The main difference between an ordinary DL subframe and the DwPTS is the number of control OFDM symbols. While the DwPTS has two control OFDM symbols; an ordinary DL subframe would have three symbols. This difference is because of the location of the primary synchronization signal (P-SCH), which is located at the third OFDM symbol in

Table 10.1 The different downlink-uplink frame configurations defined in the standard. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited

UL-DL	DL to UL switch	Subframe number									
configuration	periodicity	0	1	2	3	4	5	6	7	8	9
0	5 ms	D	S	U	U	U	D	S	U	U	U
1	5 ms	D	S	U	U	D	D	S	U	U	D
2	5 ms	D	S	U	D	D	D	S	U	D	D
3	10 ms	D	S	U	U	U	D	D	D	D	D
4	10 ms	D	S	U	U	D	D	D	D	D	D
5	10 ms	D	S	U	D	D	D	D	D	D	D
6	5 ms	D	S	U	U	U	D	S	U	U	D

the DwPTS. This difference in location enables the UEs to detect the type of duplexing implemented at the cell during network entry.

The location of the synchronization signal in FDD is located at the middle of subframe 0 and subframe 5. The GP is reserved for downlink to uplink transition.

The standard defines periodicities for frame switch-points, that is, switching from downlink to uplink and vice versa, that take place at 5 ms and 10 ms intervals. In the case of the 5 ms switch-point periodicity, a special subframe is used in every half, while in the 10 ms case, a special subframe is only used in the first half.

Table 10.1 shows the different downlink-uplink frame configurations defined in the standard. In the table, D and U are respectively downlink and uplink transmissions, while S is a special subframe for a guard time. Note the subframe 1 in all configurations and subframe 6 in configurations 0, 1, 2 and 4 (i.e., those with 5 ms switch-point periodicity). A switch-point consists of a DwPTS, GP and an UpPTS. In configurations with 10 ms switch-point periodicity, the sixth subframe is only a DwPTS. Subframes immediately following the special subframe (i.e., subframe two in all configurations and subframe seven in 5 ms periodicity) are always reserved for the UL transmission.

10.1.1 Resource Block Structure

The standard defines a resource element as the smallest time-frequency resource that can be allocated over the air. A single resource element consists of one subcarrier over one OFDMA symbol. Transmission in LTE is allocated in blocks of resource elements. A scheduler at the eNB allocates resources in Resource Blocks (RBs). Whether in UL or DL, or under FDD or TDD, a RB is 180 kHz accessed over a single time slot, that is, 0.5 ms. Alternatively, an RB can be seen as 12 contiguous subcarriers and either six or seven ODFM symbols, depending

on whether the normal or extended cyclic prefix is employed. This setup is shown in Figure 10.3. The OFDMA subcarriers spacing is 15 kHz. Depending on the implemented channel bandwidth, the number of RBs varies between 6 and 100. Specifically, for 1.4, 3, 5, 10, 15 and 20 MHz channel bandwidths, the number of RBSs is respectively 6, 15, 25, 50, 75 and 100.

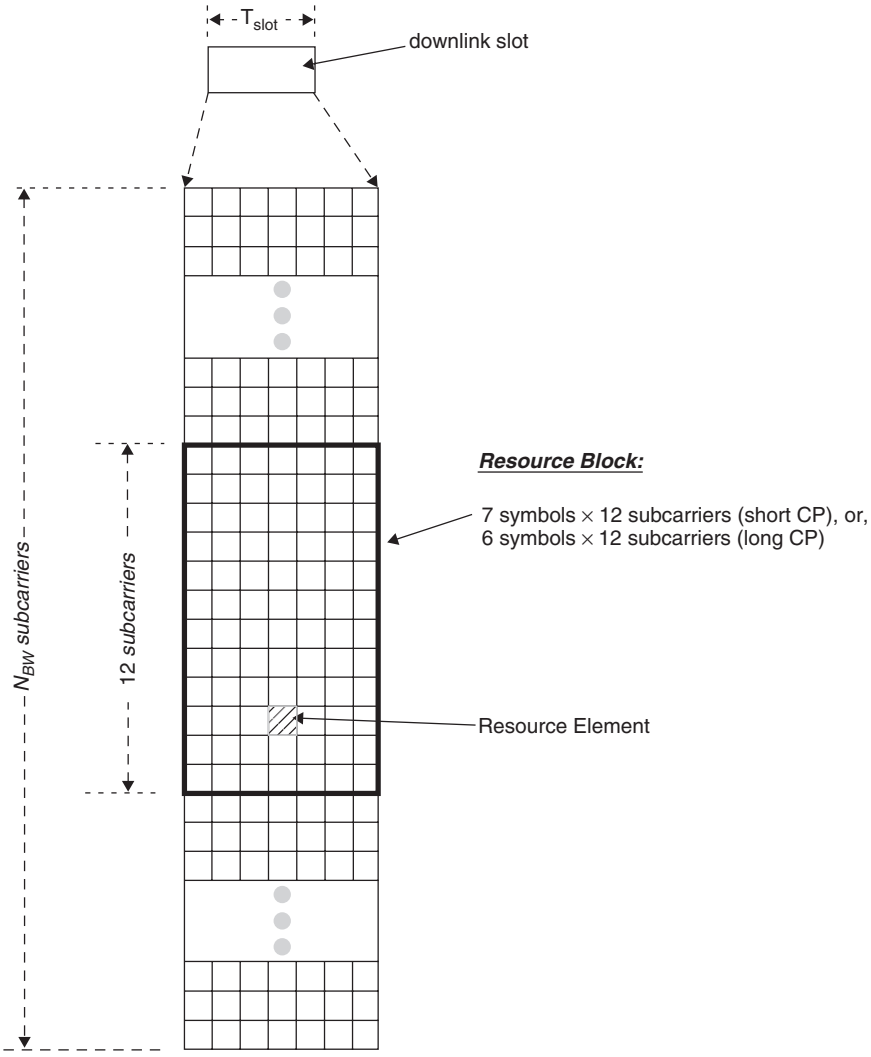


Figure 10.3 The LTE Frame. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

An UL/DL Transmission Time Interval (TTI) is one subframe in length, that is, 1 ms.

10.2 Frame-Structure in LTE-Advanced

LTE-Advanced has the same frame structure for both single-hop and relay based networks. However, the frame structure for relaying will accommodate the resource allocations of two hops, either in a centralized or distributed manner. Currently, there is no standardization for the relay frame structure.

Figure 10.4 shows a frame structure for relay-based LTE-Advanced that is similar to that of single hop with a slight difference. In Figure 10.4, the UL and the DL subframes are further organized into access and relay zones. This structure supports backwards compatibility, in addition to the communication over relay and access links. While the relay zone is used for the communication between the relay node and the eNB, the access zone is used for direct communication between the MSs and eNB or the MS and the relay node. The relay mode in LTE-Advanced is currently at its pre-draft stage. Hence, there is no detailed standardized description for the difference, if any, between Types 1 and 2 frame structure.

10.3 LTE Identification, Naming and Addressing

LTE provides detailed identification and naming values for various entities, that is, users MSs, eNBs, service areas, etc. In the following, we briefly summarize these LTE functionalities.

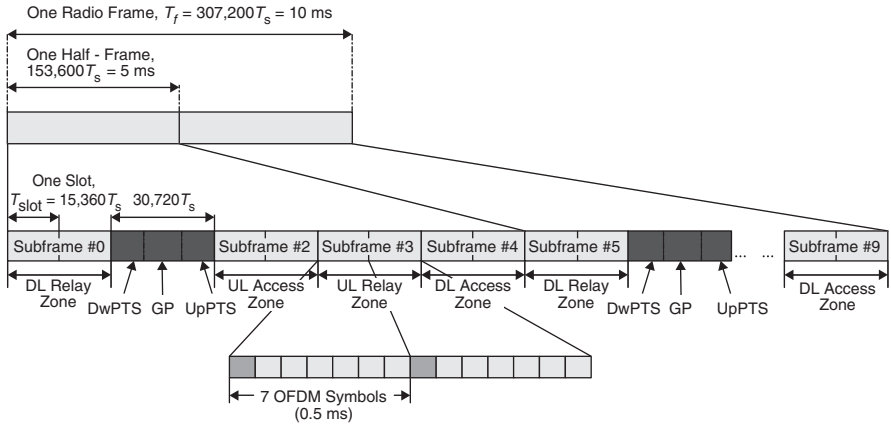


Figure 10.4 TDD-LTE-Relay frame structure: change to make it similar to the single hop relay structure forma. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

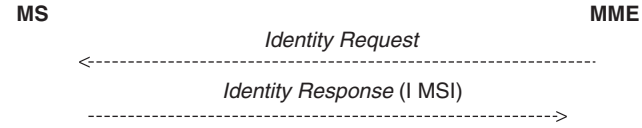


Figure 10.5 UE unique identification request and response. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

10.3.1 Identification

10.3.1.1 Identification of Mobile Subscribers

To identify mobile subscribers, LTE assigns a unique identification called the International Mobile Subscriber Identity (IMSI). The Mobility Management Entity (MME) assigns a Temporary Mobile Subscriber Identity (TMSI) to each visiting mobile subscriber in order to maintain the subscriber's confidentiality. A TMSI assigned during the network access is called the Globally Unique Temporary Identity (GUTI). The home MME correlates the allocated TMSI with the IMSI allocated to the mobile subscriber. If the GUTI is not available to provide unique identification, the serving MME requests the IMSI of the UE using a non-access-straturm procedure. This procedure is shown in Figure 10.5.

10.3.1.2 Identification of a Cell

A cell in LTE is identified by a Cell Identifier (CI) of a fixed length of 2 bytes. To generate the Cell Global Identification (CGI), the CI is concatenated with location area identification. The cell identity is unique within a location area.

10.3.1.3 Identification of Mobile Station Equipment

A mobile station equipment is uniquely and internationally identified by the fourteen decimal digits long International Mobile Station Equipment Identifier (IMEI). This identifier consists of three elements: the eight decimal digits Type Allocation Code (TAC), the six decimal digits serial number, which uniquely identifies each equipment within the TAC, and the Spare digit, which is normally set to zero when the IMEI is transmitted by the MS. Whenever the GUTI is not available, the serving MME may request the IMEI from the MS using the aforementioned procedure for requesting the IMSI.

10.3.1.4 Identification of PLMN, RNC and CN Domain

To ensure backward compatibility, LTE supports identification of RNC, CN Domain and PLMN. A PLMN is uniquely identified by its PLMN identifier (PLMN-Id) which consists of Mobile Country Code (MCC) and Mobile Network

Code (MNC). The MCC and MNC are predefined within a UTRAN while the RNC and the CN identifiers are allocated by the operator. Together, the PLMN-Id and the RNC-Id globally identify an RNC. Similarly, the CN-Id he PLMN-Id globally identify a CN node.

10.3.2 Addressing

10.3.2.1 IP Addressing

When an MS accesses a network, it can be assigned an IP address during the default bearer establishment, which in turn is maintained as long as the MS is associated with the Packet Data Network (PDN). An MS therefore retains an always-on IP connectivity with that PDN. LTE facilitates a single MS to simultaneously have multiple PDN connections. An MS will therefore establish a default bearer with each PDN, and be assigned an IP address that identifies the MS with that PDN. This facilities maintaining complete logical separation of data across the multiple networks with which the MS has simultaneous IP connectivity.

LTE provides full support for both IPv4 and IPv6 addressing. An MS can obtain an IP address as part of the MS attachment procedure (in case of default bearer establishment with IP assignment) or the IP address assignment can be obtained via DHCP or IPv6 address autoconfiguration procedure (in case of default bearer establishment without IP assignment).

10.3.2.2 MAC Addressing

Access Point Name (APN)

The APN is used as a reference to the GGSN in the GPRS backbone. The APN is translated into the IP address of the GGSN by the internal GPRS DNS to support inter-PLMN roaming. The APN is composed of two parts, the mandatory APN Network Identifier and the optional APN Operator Identifier. The APN Network Identifier defines the external network to which the GGSN is connected, while the APN Operator Identifier defines the PLMN GPRS backbone to which the GGSN is located.

HeNB Name

HeNB Name is a broadcast string of text format, which is human readable name for the Home eNB identity. The maximum length of HeNB Name is 48 bytes.

Table 10.2 provides summary of some of the identifications and naming used in LTE. For more information the reader is advised to consult 3GPP TS 23.003 document.

Table 10.2 Summary of identifications and naming used in LTE. Reproduced by permission of © 2008 3GPP. Further use is strictly prohibited

UE and Network Identities					
Name	Allocated by which SAE/ LTE entity	Purpose	Scope	Used by	Comment
I MSI	N/A	Permanent Identity of the Subscriber	Globally unique	UE, Evolved Packet Core and NAS layer (and possibly RRC during initial attach and to determine paging occasion)	For security reasons the I MSI should not be used/stored in the LTE RAN (rare exceptions might be possible)
IMEI	N/A	Permanent Identity of the end user equipment	Globally unique	UE, Evolved Packet Core and NAS layer	For security reasons the IMEI should not be used/stored in the LTE RAN (rare exceptions might be possible)
S-T MSI	MME	Temporary user identity	Unique within a tracking area or within MME pool area(s)	UE, Evolved Packet Core, NAS layer	The S-T MSI is similar to P-T MSI used today in GSM/UMTS. It needs not to share T MSI space with CS domain as in GSM/UMTS Either part of the S-T MSI is used to identify the MME within a pool of MMEs or a separate MME-ID (similar to the 3G NRI) is used (FFS)

					Solutions are needed to support network sharing Uniqueness of S-T MSI within areas depends on tracking area concept, which is FFS Format influences re-use of resolution of temporary addresses be 2G/3G SGSNs The need to store S-T MSI on legacy UICC is FFS
MS-ISDN	N/A	Permanent subscriber Identity primarily used by CS and S MS services and by O+M systems	Globally unique	UE, Evolved Packet Core, HSS, O+M	Used to identify a subscriber within a charging record. Its usage is FFS; another identifier might be used in SAE. Usage of ms-ISDN for VCC in Evolved Packet Core is FFS
IP address	IASA	Permanent or temporary Identifier used to identify the UE/user within the PDN	Unique within PDN	UE, Evolved Packet Core, PDN, operator services/I MS	The UE needs a routeable address when connected to a PDN. The IP address may be a IPv4 or a IPv6 address, and may be a private or public IP address There are also other IP address allocation options, for example, by auto-configuration

(*continued overleaf*)

Table 10.2 (*continued*)

UE and Network Identities					
Name	Allocated by which SAE/ LTE entity	Purpose	Scope	Used by	Comment
Tracking Area Identity	N/A	Permanent Identity used to identify tracking areas	Unique within a PLMN	Evolved Packet Core, UE. The tracking area identity is also broadcasted transparently in the LTE RAN	May share some similarities with the existing Routing Area Identity. Solutions are needed to support network sharing Format influences re-use of resolution of temporary addresses be 2G/3G SGSNs The need to store TAI on legacy UICC is FFS
MME Identity	N/A	Permanent Identity used to identify MME	Unique within a PLMN	Evolved Packet Core, LTE RAN, UE (indirectly via S-T MSI and Tracking Area Identity (FFS))	FFS whether a separate MME Identity is needed. As today, the old Tracking Area Identity + (parts of) the S-T MSI can identify the MME (FFS) In the LTE RAN the eNode B can (as in the RNC today) use (part of) the S-T MSI identify the MME (FFS) An MME is associated with one or more TNL, for example, IP, addresses

Cell Identity	N/A	Permanent Identity used to identify the Cell	FFS: Unique within a PLMN	Evolved Packet Core, LTE RAN	Needed to be known in the CN for some UEs in active mode when location-based charging is used [For paging, the MME needs to know which S1 interfaces to send the page message to. Hence the MME probably needs to be able to map Tracking Area to Cell IDs/eNodeB IDs. Note that the cells within an eNodeB may need to be in different tracking areas.] It is FFS if the Cell Identity is associated with a TNL address FFS whether Cell Identity has to be unique within LTE RAN or globally unique or whether both are needed
eNode B Identity	N/A	Permanent Identity used to identify the eNode B	Unique within a PLMN	Evolved Packet Core, LTE RAN	FFS whether a specific eNodeB identity is needed or whether (TNL) addresses are sufficient. Used to derive (TNL) addresses for S1 addressing The eNode B Identity is associated with one or more TNL, for example, IP, addresses

(*continued overleaf*)

Table 10.2 (*continued*)

UE and Network Identities

Name	Allocated by which SAE/LTE entity	Purpose	Scope	Used by	Comment
eNode B Specific S1 UE Context Identity	eNode B	Temporary identity used to identify an S1 UE context within eNodeB	FFS: Unique within a eNode B [and x2 interface handover target eNodeBs?]	Evolved Packet Core, LTE RAN	It is used to identify the MME and/or UPE UE context(s) in the eNode B that relate to signaling relations(s) over S1 Whether MME, UPE or both UE context identities are needed depends on function separation between MME and UPE. NAS signaling, for example, might be exchanged over S1 by using UE/user identity without a need for an additional UE Context Identity The S1 addressing principle is still FFS Uniqueness within a (TNL) eNodeB address might be sufficient
MME Specific S1 UE Context Identity	MME	Temporary identity used to identify an S1 UE context within MME	FFS: Unique within a MME	Evolved Packet Core, LTE RAN	It is used for signaling over S1 to identify the UE context in the MME. Its need is FFS The S1 addressing principle is still FFS

					Uniqueness within a (TNL) MME address might be sufficient
UPE Specific S1 UE Context Identity	UPE	Temporary identity used to identify an S1 UE context within UPE	FFS: Unique within a UPE	Evolved Packet Core, LTE RAN	It is used for signaling over S1 to identify the UE context in the UPE. Its need is FFS The S1 addressing principle is still FFS Uniqueness within a (TNL) UPE address might be sufficient
UPE identity	N/A	Permanent Identity used to identify the UPE from the LTE RAN	Unique within a PLMN	Evolved Packet Core, LTE RAN	FFS if needed or if the UPE TNL address, for example, IP address, is enough
PDN Identity	N/A	Permanent Identity used to identify one or multiple specific PDN(s)	Globally unique	UE, Evolved Packet Core	Depending on Multiple PDNs solution. It may be an APN
PCRF Id	N/A	Permanent Identity used to identify the PCRF	Unique within a PLMN	Evolved Packet Core	FFS if needed or if the PCRF TNL address is enough
HSS Id	N/A	Permanent Identity used to identify the HSS	Unique within a PLMN	Evolved Packet Core	FFS if needed or if the HSS is identified by (part of) user Identities

(*continued overleaf*)

Table 10.2 (*continued*)

UE and Network Identities					
Name	Allocated by which SAE/LTE entity	Purpose	Scope	Used by	Comment
RAT ID	N/A	Radio Access Technology Identity used to identify the type of radio access technology	Globally unique	Evolved Packet Core	FFS if needed or is used by the PCC of SAE

11

UE States and State Transitions

The 3GPP standard defines simplified states and state transitions for LTE UE. This simplification is relative to previous standards such as UTRAN and GSM. At any given time, a powered on LTE UE is either IDLE or CONNECTED. LTE-Advanced relies on the same state definitions, and the following descriptions apply for both LTE and LTE-Advanced.

This chapter is organized as follows. Section 11.1 provides an overview of the UE state definitions, and the transitions between the different states. Section 11.2 is dedicated to describing processes performed when a UE is in the IDLE state, including processes for PLMN selection, cell selection and reselection, location registration, and support for manual CSG ID selection. In Section 11.3, the procedures for acquiring system information that are required for cell or network selections are described. Section 11.4 goes over the procedures for establishing and control connections in LTE and LTE-Advanced. These procedures involve the random access mechanisms exercised to gain initial access to the network, in addition to dedicated signaling for connection establishment and re-establishment, connection reconfiguration, in addition to process engaged when a UE leaves the connected state. The final section in the chapter, Section 11.5, describes how the access stratum states described throughout are mapped to the NAS states.

11.1 Overview of a UE's State Transitions

Interaction between the AS and the NAS is maintained as long as a UE is powered on. As previously described, a UE's communication with the RAN is performed over the AS, while communication with a network's core is performed through the NAS.

When powered on, a UE first searches for a Public Land Mobile Network (PLMN). If successful, the UE then attempts to camp on a cell in the PLMN. This is performed through a cell search, which depends on frequency and timing synchronization between the UE and the selected eNB. The standard differentiates

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

between camping on a cell and camping on any cell, where in the latter the UE camps regardless of the PLMN identity. Once camped, a UE is ready to initiate connections, or to maintain connectivity while mobile. To be connected, the UE goes through a connection setup that relies on RRC for setup, (re)configuration and security. Once a connection terminates, it is released.

Note that a UE can camp on either an acceptable or a suitable cell. An acceptable cell is one that allows the UE to initiate emergency calls, and to receive ETWS or CMAS can be received. A suitable cell, on the other hand, is one where the UE can achieve full connectivity, depending on its capabilities. In case of a CSG HeNB, the cell would be on the UE's white list, that is, among the cells accessible by the UE.

Resources for this chapter can be sought as follows.

- Overall description of RRC services, states and transitions can be found in 36.300.
- Detailed descriptions for IDLE processes are described in 36.304 (AS) and 23.122 (NAS).
- PHY layer aspects of cell search are described in 36.213.
- Descriptions for RRC processes in both IDLE and CONNECTED states are described in 36.331.

11.2 IDLE Processes

In EUTRAN, there are four processes in the idle mode, as schematized in Figure 11.1.

- PLMN selection
- Cell selection and reselection
- Location registration
- Support for manual CSG ID selection

The following describes these processes in detail.

11.2.1 PLMN Selection

PLMN selection is made by the UE's NAS through relaying the identifier of the selected PLMN. To do so, the UE requires knowledge of identified PLMNs and supporting measurements from the AS. This relay of identified PLMNs requires the UE's synchronization to a broadcast channel, and can be either on request or autonomously. PLMN selection can be done either automatically (without user interruption) or manually. The AS, in its search and measurements, identifies to the NAS PLMNs with both high and low quality signals. An optimization for PLMN search can be made through using stored information, such as carrier frequencies. At any instant, the search may be stopped on request from the NAS. Once completed, the UE proceeds to the cell selection process.

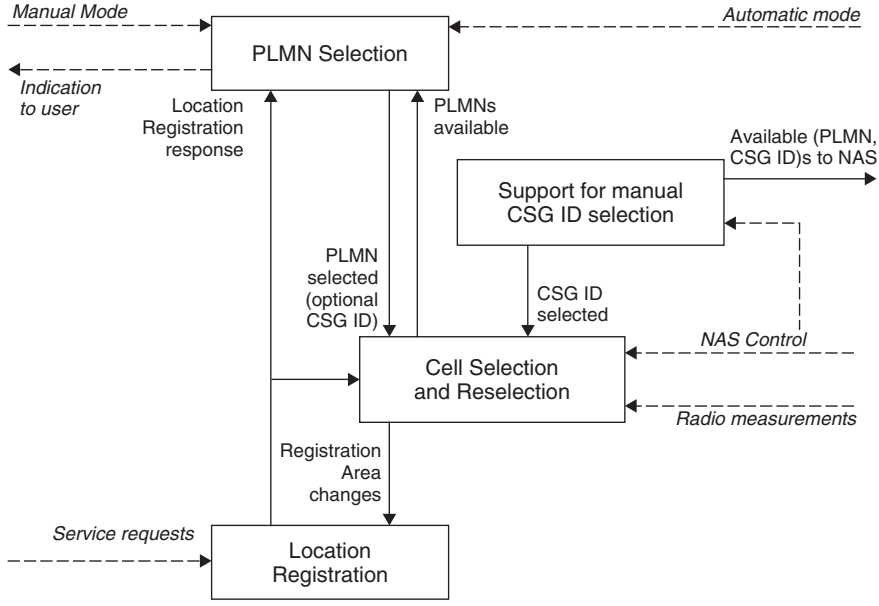


Figure 11.1 IDLE Mode Process. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

If a CSG ID is passed along to the NAS during the PLMN search process, the UE must consider either an acceptable or a suitable cell belonging to the provided CSG ID to camp on. The AS will inform the NAS if the UE is no longer camped on a cell with the provided CSG ID.

11.2.2 Cell Selection and Reselection

The objective of the cell selection procedure is to identify a cell on which the UE can camp to receive either limited or normal services. Limited services are received when a UE is camped on an acceptable cell, while normal services are received when camped on a suitable cell.

Cell selection starts with performing measurements to support the selection decision. The details of such measurements are described in 36.133. In the AS, the UE then goes into detecting and synchronizing to a broadcast channel, receives and handles broadcast information, and forwards NAS system information to the NAS. Meanwhile, the NAS controls the cell selection process, for example, through indicating the RAT(s) associated with the selected PLMN to be used in initial cell selection. The NAS also maintains lists of forbidden registration areas and whitelist CSG IDs. In turn, the AS directs the UE's search to the indicated RAT(s). If the NAS also provides a CSG whitelist, the AS verifies whether the CSG is suitable.

In LTE, there are two signals transmitted in the downlink to facilitate the cell search, the primary synchronization signal and the secondary synchronization. Whether or not the UE is in the energy saving Discontinuous Reception (DRX) mode, the UE will monitor radio frames to asses radio link quality and recognize whether it in- or out-of-sync. The UE will report to higher layers its recognition of both signal quality and synchronization status.

If while IDLE the UE finds a more suitable cell to camp on, the NAS is informed of a reselection. This can be due to strong variation in a selected cell's signal quality, or due to the UE's mobility. A cell may also trigger a reselection, for example, for load balancing or when reselection evaluation procedures are changed. The standard specifies proper triggers and frequencies for undergoing cell reselection.

11.2.3 Location Registration

Location registration procedures are performed at the NAS. The AS merely transfers registration area information relayed by the eNB to the NAS. In turn, the NAS oversees procedures for initial location registration, registration update (e.g., when entering a new tracking area), maintenance of forbidden registration areas, and deregistration when a UE is properly shut down.

11.2.4 Support for Manual CSG ID Selection

Additional functionalities are provided between the AS and the NAS to support Manual CSG ID Selection. This includes the AS searching for cells with a CSG ID, reading the HNB names and selecting a CSG ID based on criteria provided by the NAS.

11.3 Acquiring System Information

While IDLE, a UE needs to acquire system information prior to getting connected. This process of system information acquisition also takes place at other instances, and in both IDLE and CONNECTED states. Such instances include reselection, recovery from link failure, both intra- and inter-RAT handoffs, and when the UE receives notification of change in system information.

System information is divided into information blocks. The standard identifies a MasterInformationBlock (MIB) and several types SystemInformationBlocks (SIBs). The MIB defines the most essential physical layer information of a cell required to receive further system information. Type 1 SIB (i.e., SystemInformationBlockTYpe1) contains information relevant when evaluating if a UE is allowed to access a cell and defines the scheduling of other SIBs. Type 2 SIB contains common and shared channel information, while Type 3 SIB contains cell re-selection information, mainly related to the serving cell. Other types of SIB carry are used to relay other specific information such availability of

other EUTRA frequencies, neighboring RATs (UTRA, GERAN, CDMA2000), HNBID, ETWS, CMAS and MBMS-related information.

The MIB is transferred over the BCH. SIBs other than Type 1 SIB are carried in SystemInformation (SI) messages. Type 1 SIB also carries scheduling information for other SIB types. All SIB are transmitted on DL-SCH. Note, however, that MIB and SIB follow a fixed schedule, MIB with a periodicity of 40 ms and Type 1 SIB of 80 ms.

Apart from ETWS and CMAS, change of system information occurs per a concept of a modification period. Within a modification period system information may be transmitted a number of times with the same content. When some of or all of the system information changes, the network notifies the UE any time during a modification period, and the updated system information is transmitted in the next modification period. Until notified of and acquired new system information, a UE applies the system information it has already acquired.

The TS 36.331 describes in detail the UE's response to receiving the MIB and various SIBs. For example, upon receiving a MIB, the UE shall apply the MIB's resource configuration. Whether IDLE or connected, it should also apply the received values of downlink and uplink bandwidth allocations until SIB Type 2 is received. When Receiving type 1 SIB, the UE considers whether it supports the indicated frequency bands. If the UE does not, it would consider the cell barred. Otherwise, the UE would forward received cell identity and tracking area code for upper layers. A UE cannot initiate an RRC connection establishment procedure until it has a valid version of MIB, and SIB types 1 and 2.

11.4 Connection Establishment and Control

Prior to describing RRC procedures for connection establishment and, in general, control, a description of Signaling Radio Bearers (SRB) is required. SRBs are RBs that are used solely for the transmission of RRC and NAS messages. SRB0 is used for RRC messaging using the CCCH logical channel, which is the channel used for UEs having no RRC connection with the network. On the other hand, SRB1 utilizes DCCH and is used for RRC messages (including piggybacked NAS messages), as well as NAS messages transmitted prior to establishing SRB2. Note that the establishment of an RRC connection involves the establishment of an SRB1. SRB2 also utilizes DCCH and is used for NAS messages. SRB2 has a lower priority than SRB1, and is always configured after security activation.

11.4.1 Random Access Procedure

Connection establishment begins with a procedure for random access through which the user is scheduled a transmission to initiate connection setup. There are other instances where the random access procedure may be required. These include RRC connection re-establishing, handover, upon data arrival in the uplink or downlink while the UE is not synchronized or allocated resources, and for positioning while in the CONNECTED state.

There are two types of random access procedure: contention based, and non-contention based. The non-contention based procedure is utilized when it is the network that is trying to establish the connection. The contention based procedure is hence mostly utilized for UE initiated activity.

The contention based procedure, shown in Figure 11.2 involves the following.

1. The user first generates and sends a Random Access Preamble on the RACH in the uplink.
2. The eNB responded with a Random Access Response on the DL-SCH. This response conveys at least the preamble identifier, Timing Alignment information, initial UL grant and the assignment of Temporary C-RNT1.
3. The UE then, when scheduled, transmits its scheduled transmission on the UL-SCH. This scheduled response varies depending on whether the UE is attempting initial access, connection re-establishment, after handover (in the target cell) or other reasons.
4. Finally, indication of contention resolution is made on PDCCH by the eNB. This is indication is not synchronized with the UE's scheduled transmission, and is addressed to either the Temporary C-RNTI or C-RNTI depending on whether the UE is attempting initial access or is CONNECTED.

It should be noted that in the contention-based procedure that L1 only oversees the preamble-response exchange, and that the remaining exchanges are scheduled and overseen by higher layers.

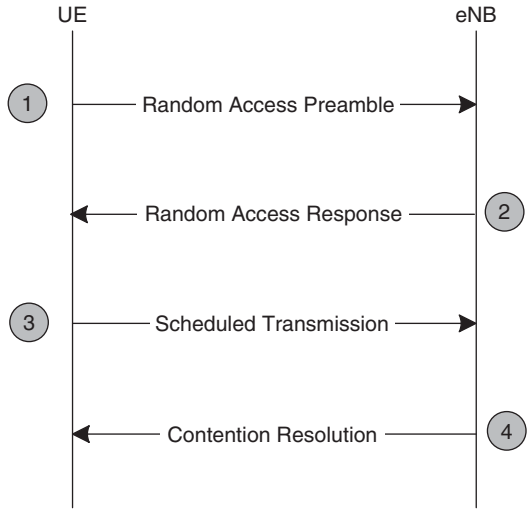


Figure 11.2 Contention based Random Access Procedure. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

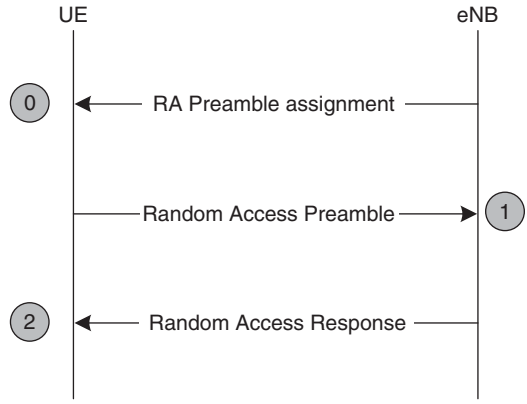


Figure 11.3 Non-contention based Random Access Procedure. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

The non-contention based random access procedure, shown in Figure 11.3 involves the following steps:

1. An assignment of a RA preamble by the eNB to the UE made through either a HO command or PDCCH in case of DL data arrival.
2. The UE then sends the RA preamble on the RACH in the uplink.
3. The eNB then sends a RA response on the DL-SCH.

11.4.2 Connection Establishment

A successful RRC connection establishment exchange is shown in Figure 11.4.

A UE initiates the connection establishment through sending an RRCConnectionRequest. This request is scheduled through a contention based random access procedure, as described above. However, prior to attempting to establishing an RRC connection, the UE verifies whether or not the cell it is camped on is barred or not. This verification is made through processing the cell's SIBs. Note that a UE is not required to ensure that it maintains updated system information applicable only for UEs in RRC_IDLE state. However, the UE needs to perform system information acquisition upon cell re-selection. The UE shall also, while awaiting the eNB's response, continue cell re-selection related measurements and evaluation and, if conditions for re-selections are fulfilled, perform cell re-selection.

The EUTRAN responds to an RRCConnectionRequest with an RRCConnectionSetup. A UE receiving the RRCConnectionSetup applies the relayed configurations, enters the RRC_CONNECTED state, and stops the cell re-selection procedure. The UE then responds with an RRCConnectionSeteup

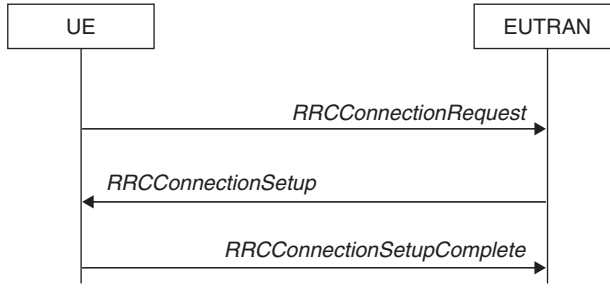


Figure 11.4 A successful RRC Connection Establishment. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

complete message carrying PLMN and MME (if registered) related information, in addition to NAS information received from upper layers.

If upper layers abort the RRC connection establishment procedure before the UE enters the RRC_Connected state, the UE resets the MAC, releases MAC configuration and re-establishing RLC for all established RBs.

If the EUTRAN rejects the UE's RRCConnectionRequest, it responds with an RRCConnectionReject. Upon receiving a reject messages, the UE releases the MAC configuration and indicates the rejection to the higher layers.

11.4.3 Connection Reconfiguration

LTE allows for connection reconfiguration in instances where its required to established, modify or release RB, undergo handover, or to setup or release measurements. A successful connection reconfiguration is shown in Figure 11.5. The exchange involves the EUTRAN sending an RRCConnectionReconfiguration and the UE responding with an RRCConnectionReconfigurationComplete. Note that EUTRAN cannot relay the control information to the UE, nor establish RBs other than SRB1 if the AS security has not been activated. For example, AS

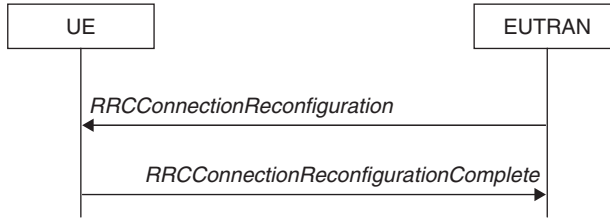


Figure 11.5 A successful RRC Connection Reconfiguration. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

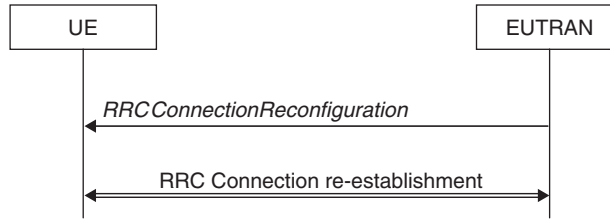


Figure 11.6 A failure RRC Connection Reconfiguration. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

security needs to be activated before the EUTRAN sends mobilityControlInfo for handover purposes.

If a UE is unable to apply all or part of the reconfiguration, the connection reconfiguration is considered a failure, as is shown in Figure 11.6. The UE will continue using prior configuration and initiate connection re-establishment, which will be described in the following section. If AS security has not been activated, the UE begins procedures to leave the RRC_CONNECTED state.

11.4.4 Connection Re-establishment

Connection re-establishment involves the resumption of SRB1 operation and the reactivation of AS security. A UE in RRC_CONNECTED state with AS security activated may initiate re-establishment with a cell that has a valid UE context. If EUTRAN accepts re-establishment, SRB1 operation resumes but the operation of other RBs remains suspended.

Re-establishment is motivated by various failures, including failures in radio link, handover, mobility, integrity check or RRC connection reconfiguration. In initiating the request, the UE suspends all RBs except SRB0, resets MAC configurations, applies default configuration based on most up to date system configuration and performs cell selection.

If the EUTRAN rejects the re-establishment requests, the UE leaves the RRC_-CONNECTED state and performs relevant actions.

11.4.5 Connection Release

A EUTRAN releasing the connection of a UE in RRC_CONNECTED state sends an RRCConnectionRelease message to the UE. In response, the UE performs actions relevant to leaving the RRC_CONNECTED state and applies idle mode mobility control info indicated in the release message.

A UE may also respond to a connection release requested by upper release. In such a case, the UE performs actions made when leaving the RRC_CON-NECTED.

11.4.6 Leaving the RRC_CONNECTED State

When leaving the RRC_CONNECTED state, the UE resets the MAC; releases all radio resources, including the release of the RLC entity, MAC configuration and the associated PDCP for all established RBs; indicates the connection release together with the release cause to the upper layers; and perform cell selection.

In cell selection, the UE will select a suitable cell to camp on based the most recent available system information, for example, if indicated in the RRCConnectionRelease message.

11.5 Mapping between AS and NAS States

In EUTRAN, a UE transitions between RRC_IDLE and RRC_CONNECTED. The UE's state model in view of the NAS is two dimensional. The first dimension defines the EPS Mobility Management (EMM) states that dictate whether a UE is attached or detached, and the tracking area to which the UE belongs. The second dimension details the EPS Connection Management (ECM) states that essentially describe the connectivity between the UE and the network's core.

The two relevant EMM states are EMM-DERGSTERED and EMM-REGISTERED. The two relevant ECM states are ECM-IDLE and ECM-CONNECTED. While the EMM and ECM states are independent, for example, a UE can become (EMM)-deregistered regardless of the ECM state, there are certain relations between the two, and between both NAS states and the AS states.

An EMM-DERIGSTERED UE is one with valid location or routing information and is hence unreachable by the MME. An EMM-REGISTERED UE, on the other hand, is one that has been attached to either a EUTRAN or GERAN/UTRAN network. In such a state, the UE's location in the MME is known to at least an accuracy of the tracking area. A UE returns to the DERIGESTERED state upon a detach procedure or upon being handed over to a non-3GPP network.

An ECM-IDLE state indicates that no connection for NAS signaling has been setup between the UE and the core. An UE in ECM-IDLE needs to perform PLMN and cell selection and reselection in order to become ECM-CONNECTED. Once ECM-CONNECTED, a UE location is known in the MME by the serving eNB ID and the UE's mobility is handled by the handover procedures.

The relation between the NAS and AS states is as follows:

- A UE that is EMM-DERIGSTERED and ECM-IDLE is in RRC_IDLE. Mobility management in this case entails PLMN selection, and the UE's position is not known by the network.

- A UE that is EMM-REGISTERED and ECM-IDLE is also in RRC_IDLE. Mobility management here entails cell reselection, and the UE's position at tracking area level.
- A UE that is EMM-REGISTERED AND ECM-CONNECTED with radio bearers established is in RRC_CONNECTED. Handover procedures (with RRC, X1 or S2 signaling) handle mobility management, and UE's position is known by the network at the cell level.

12

Quality of Service and Bandwidth Reservation

This chapter discusses how quality of service management and bandwidth provisioning are performed in LTE (based on 3GPP Release 9) and LTE-Advanced (based on Release 10). The standard is presented in a manner that illustrates measures for QoS performance, service classification, signaling for bandwidth requests and grants, and bandwidth allocation and traffic handling.

The chapter is organized as follows. Section 12.1 introduces 3GPP's measures of QoS performance, while Section 12.2 discusses traffic classification. Section 12.3 Reviews the signaling for making bandwidth requests and grants, and describes the distinctions between the dedicated bearer and the default bearer. Meanwhile, Section 12.4 discusses bandwidth allocation and traffic handling, and offers an overview of LTE scheduling, both in the OFDMA downlink and the SC-FDMA uplink, in addition to the wireless channel reliability mechanisms. QoS in LTE-Advanced is described in Section 12.5, and provides descriptions for the IMT-Advanced technology's new features such as carrier aggregation, coordinated multipoint transmission and relaying.

12.1 QoS Performance Measures

To achieve a QoS for a certain application, the application requirements must be quantified in terms of parameters that identify the target performance level. Such a level is normally measured in terms of throughput, delay, jitter, and packet loss.

LTE identifies the following major quantitative parameters.

1. *Throughput*: Characterized through the Guaranteed Bit Rate, Maximum Bit Rate and Aggregate Maximum Bit Rate.

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

a) *The Guaranteed Bit Rate (GBR)*: Network resources allocated based on GBR are fixed and do not change after bearer establishment or modification. This is hence a guaranteed service data flow.
b) *The Maximum Bit Rate (MBR)*: This parameter limits the bit rate that can be expected to be provided to GBR bearer, and is enforced by network shaper to restrict the traffic to its maximum bit rate agreement.
c) *Aggregate Maximum Bit Rate (AMBR)*: This parameter is used for non-GBR flows, and has two types, APN-AMBR and UE-AMBR. The APN-AMBR (Access Point Name-AMBR) is a subscription parameter stored at the HSS per APN. The HSS defines a QCI for each PDN (identifiable by an individual PDN identifier) and an APN-AMBR for each ARP. The APN-AMBR parameter refers to the maximum bit rate that can be consumed by all non-GBR bearers and all PDN connections of this APN. This parameter is enforced by P-GW in the downlink and by both UE and P-GW in the uplink. The UE-AMBR parameter, on the other hand, refers to the maximum bit rate allowed for all non-GBR bearer aggregates for the respective UE. The parameter is enforced in both the downlink and the uplink.

Note that GBR and MBR are defined per bearer while the AMBR parameters are defined per a group of bearers. All throughput parameters have two components, one for downlink and another for uplink.

2. *Delay*: Specified by the packet delay budget. LTE defines nine categories for delay, with 50 ms being tightest and 300 ms being the slackest. The latter value is used for delay tolerant applications.
3. *Packet Loss*: Defined as the Packet Error Loss Rate, and is similar to the packet delay budget in having nine categories with 10^{-6} being best and 10^{-2} being the worst.
4. *Priority*: Specified by the Allocation/Retention Priority (APR) parameter, which is used to indicate the priority of both allocation and retention of the service data flow. The APR dictates whether a bearer establishment/modification request can be accepted or rejected in the event of conflicts in demand for network resources. At the time of exceptional network resources limitations, such as handover, ARP can be used by the eNodeB to drop a flow with a lower ARP to free up capacity. ARP, however, has no effect on the network treatment received by the flow once the flow is successfully established.

12.2 Classification

LTE classifies flows into GBR and non-GBR, LTE also differentiates between Radio Bearers, S1 Bearers and EPS bearers. A radio bearer is the over-the-air connection. An S1 bearer is the connection between the eNodeB and the MME/SGW. Finally, the EPS bearer is established between the EPS and the MME and the SGW, and the SGW and the PGW.

Table 12.1 An example of QoS classes identified by the QCI. Reproduced by permission of © 2011 3GPP. Further use is strictly prohibited

QCI	Resource Type	Priority	Packet Delay Budget	Packet Error Loss Rate	Example Services
1	GBR	2	100 ms	10^{-2}	Conversational Voice
2	GBR	4	150 ms	10^{-3}	Conversational Video (Live Streaming)
3	GBR	5	300 ms	10^{-6}	Non Conversational Video (Buffer and playback)
4	GBR	3	50 ms	10^{-3}	Real Time Gaming
5	GBR	1	100 ms	10^{-6}	I MS Signaling
6	GBR	7	150 ms	10^{-3}	Voice, Video, Interactive gaming
7	GBR	6	300 ms	10^{-6}	Video (Buffer and playback)
8	GBR	8	300 ms	10^{-6}	TCP Based
9	GBR	9	300 ms		

A default bearer is initiated and established at the startup time to carry all traffic. The default bearer is a non-GBR bearer, and does not provide bit rate guarantees. A dedicated bearer can be either a GBR or a non-GBR bearer. If a GBR, it can specify the guarantee dbit rate, packet delay and packet loss error rate. Each dedicated bearer is characterized by a TFT with QoS parameters associated to it. An uplink TFT is used to map the UE uplink traffic to specific QoS parameters, with the mapping carried out at both the eNodeB and the UE. Mapping for the downlink TFT is carried out at the SGW or the PGW. Table 12.1 gives an example of a traffic classification based on the QoS parameters defined the LTE QoS framework. Each class is identified by a scalar number called the QoS class Identifier (QCI). A QCI is identifies a group of QoS parameters describing the packet forwarding treatment in terms of priority, allowable delay, and packet error rate.

12.3 Signaling for Bandwidth Requests and Grants

The setup, management and release of a radio bearer are carried out by the MME on the S1 interface. This is an addition of other functionalities required for bandwidth requests and grants signaling such as resource management and admission control. The setup release and management of a dedicated radio bearer are different from those made for a default bearer, as will be shown in the following sections.

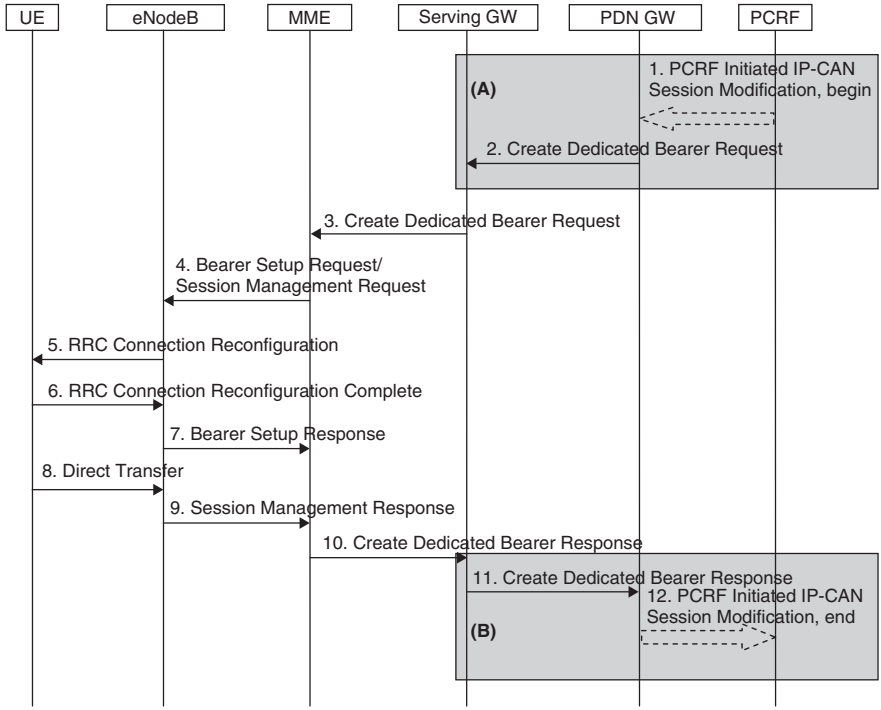


Figure 12.1 Dedicated bearer activation. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

12.3.1 Dedicated Bearer

12.3.1.1 Dedicated Bearer Activation

The activation of dedicated radio, shown in Figure 12.1, bearer takes place as follows:

1. The PDN GW sends a Create Dedicated Bearer Request message (I MSI, PTI, EPS Bearer QoS, TFT, S5/S8 TEID, LBI, Protocol Configuration Options) to the Serving GW. The LBI is the Linked EPS Bearer Identity (LBI) of the default bearer.
2. The Serving GW sends the Create Dedicated Bearer Request (I MSI, PTI, EPS Bearer QoS, TFT, S1-TEID, LBI, Protocol Configuration Options) message to the MME. If the UE is in ECM-IDLE state, the MME will trigger the Network Triggered Service Request from Step 3. In such a case, Steps 4-7 may be combined into a Network Triggered Service Request procedure or be performed individually.
3. The MME selects an EPS Bearer Identity that has not yet been assigned to a UE. The MME then builds a Session Management Request carrying the PTI,

TFT, EPS Bearer QoS parameters (excluding ARP), Protocol Configuration Options, the EPS Bearer Identity and the LBI. If the UE has UTRAN or GERAN capabilities, the MME uses the EPS bearer QoS parameters to derive the corresponding PDP context parameters QoS Negotiated (R99 QoS profile), Radio Priority, Packet Flow Id and TI and includes them in the Session Management Request. Then MME then signals the Bearer Setup Request (EPS Bearer Identity, EPS Bearer QoS, Session Management Request, S1-TEID) message to the eNodeB.

4. A Bearer Setup Request and a Session Management Request are then sent by the MME to the eNodeB.
5. The eNodeB acknowledges the bearer activation to the MME with a Bearer Setup Response (EPS Bearer Identity, S1-TEID) message. The eNodeB indicates whether the requested EPS Bearer QoS could be allocated.
6. The eNodeB sends an Uplink NAS Transport (Session Management Response) message to the MME.
7. Upon reception of the Bearer Setup Response message (Step 7) and the Session Management Response message (Step 9), the MME acknowledges the bearer activation to the Serving GW by sending a Create Dedicated Bearer Response (EPS Bearer Identity, S1-TEID) message.
8. The Serving GW acknowledges the bearer activation to the PDN GW by sending a Create Dedicated Bearer Response (EPS Bearer Identity, S5/S8-TEID) message.
9. Dedicated Bearer activated.

12.3.1.2 Bearer Deactivation

Bearer deactivation, shown in Figure 12.2, takes place as follows:

1. If dynamic PCC is not deployed, the PDN GW is triggered to initiate the Bearer Deactivation procedure due to either a change in QoS policy or based on a request from the MME. Optionally, the PCRF sends QoS policy to the PDN GW. This corresponds to the initial steps of the PCRF-initiated IP-CAN Session Modification procedure or the response to the PCEF initiated IP-CAN Session Modification procedure as defined in TS 23.203, up to the point that the PDN GW requests IP-CAN Bearer Signaling. If dynamic PCC is not deployed, the PDN GW may apply local QoS policy. The PDN GW initiated Bearer deactivation is also performed when handovers without optimization occur from 3GPP to non-3GPP, in which case the default bearer and all the dedicated bearers associated with the PDN address are released. The PDN address, however, is retained in the PDN GW.
2. The PDN GW sends a Delete Bearer Request message (PTI, EPS Bearer Identity, Causes) to the Serving GW. The Procedure Transaction Id (PTI) parameter in this and the following steps is used only when the procedure is initiated by a UE Requested Bearer Resource Modification Procedure. This message can include an indication that all bearers belonging to that

PDN connection shall be released. The PDN GW includes 'Cause' IE in the Delete Bearer Request message and sets the IE to 'RAT changed from 3GPP to Non-3GPP' if the Delete Bearer Request message is caused by handover without optimization occurs from 3GPP to non-3GPP.

3a. The Serving GW sends the Delete Bearer Request (PTI, EPS Bearer Identity, Cause) message to the MME. This message can include an indication that all bearers belonging to that PDN connection shall be released.

3b. If ISR is activated, the Serving GW sends the Delete Bearer Request (PTI, EPS Bearer Identity, Cause) message to the SGSN. This message can include an indication that all bearers belonging to that PDN connection shall be released, and the SGSN releases all bearer resources of the PDN connection.

4. If the release of the bearer in E-UTRAN has already been signaled to the MME, Steps 4–7 are omitted. Otherwise the MME sends the S1-AP Deactivate Bearer Request (EPS Bearer Identity) message to the eNodeB. The MME builds a NAS Deactivate EPS Bearer Context Request message including the EPS Bearer Identity, and includes it in the S1-AP Deactivate Bearer Request message. When the bearer deactivation procedure was originally triggered by a UE request, the NAS Deactivate EPS Bearer Context Request message includes the PTI.

5. The eNodeB sends the RRC Connection Reconfiguration message including the EPS Radio Bearer Identity to release to the UE. If the S1-AP message in step 4 contains a NAS PDU, the RRC message includes the NAS PDU.

6a. The UE RRC releases the radio bearers indicated in the RRC message in step 5, and indicates the radio bearer status to the UE NAS. Then the UE NAS removes the UL TFTs and EPS Bearer Identity according to the radio bearer status indication from the UE RRC. The UE responds to the RRC Connection Reconfiguration Complete message to the eNodeB.

6b. The eNodeB acknowledges the bearer deactivation to the MME with a Deactivate Bearer Response (EPS Bearer Identity) message.

7a. The UE NAS layer builds a Deactivate EPS Bearer Context Accept message including EPS Bearer Identity. The UE then sends a Direct Transfer (Deactivate EPS Bearer Context Accept) message to the eNodeB.

7b. The eNodeB sends an Uplink NAS Transport (Deactivate EPS Bearer Context Accept) message to the MME.

8a. The MME deletes the bearer context related to the deactivated EPS bearer and acknowledges the bearer deactivation to the Serving GW by sending a Delete Bearer Response (EPS Bearer Identity) message.

8b. The SGSN deletes PDP Context related to the deactivated EPS bearer and acknowledges the bearer deactivation to the Serving GW by sending a Delete Bearer Response (EPS Bearer Identity) message.

9. If ISR is activated, after receiving the two Delete Bearer Response messages from the MME and the SGSN, or if ISR is not activated, after receiving the Delete Bearer Response messages from the MME, the Serving GW deletes the bearer context related to the deactivated EPS bearer acknowledges the

bearer deactivation to the PDN GW by sending a Delete Bearer Response (EPS Bearer Identity) message.

10. The PDN GW deletes the bearer context related to the deactivated EPS bearer. If the dedicated bearer deactivation procedure was triggered by receiving a PCC decision message from the PCRF, the PDN GW indicates to the PCRF whether the requested PCC decision was successfully enforced by completing the PCRF-initiated IP-CAN Session Modification procedure or the PCEF initiated IP-CAN Session Modification procedure as defined in TS 23.203 [6], proceeding after the completion of IP-CAN bearer signaling.

12.3.2 Default Bearer

The default bearer setup is initiated when the terminal first connect to the PDN. It is a non-GBR bearer, which remains active as long as the UE is associated

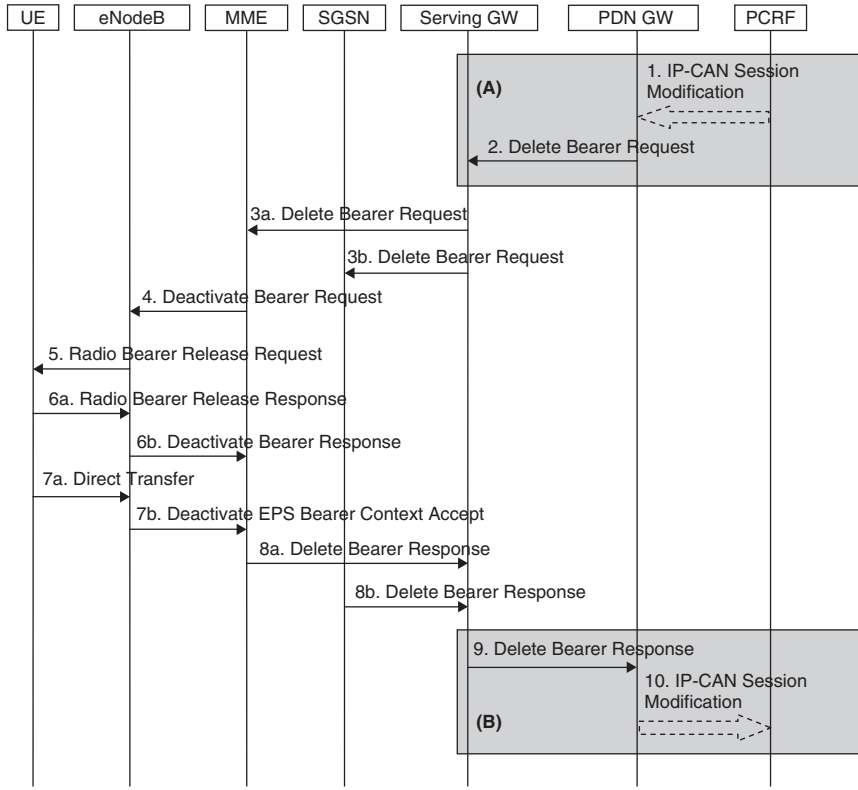


Figure 12.2 Dedicated bearer deactivation. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

with a cell. Once a node is disconnected from the cell, the default bearer is released. Hence, the main objective of the default bearer is to provide the user with ubiquitous access. More details about the default bearer connection setup, management and release can be found in Section 12.3.

12.4 Bandwidth Allocation and Traffic Handling

12.4.1 Scheduling

The scheduler resides in the eNodeB to dynamically allocate uplink and downlink resources over the uplink and downlink shard channel U-SCH and D-SCH, respectively. Uplink scheduling is performed per SC-FDMA while downlink is performed for OFDMA. The eNodeB calculates the time-frequency resources given the traffic volume and the QoS requirements of each radio bearer. However, the resources are allocated per UE and not per radio bearer.

The uplink and downlink schedulers are invoked to allocate resources every TTI. The minimum TTI duration is of one subframe length; that is, 1 ms. However, the LTE specification allows adaptive downlink TTI duration where multiple subframes can be concatenated to produce a longer TTI duration. This concatenation reduces the overhead for higher layers. The TTI length can be set dynamically by the eNodeB through defining the modulation and coding scheme used and the size of the resource blocks. Otherwise, it can be set semi-statically through higher layer signaling. Adaptive TTI length can be used to improve the Hybrid Automatic Repeat Request (HARQ) performance or the support of lower data rates and quality of service. In the following two sections we summarize the operation of the downlink scheduler and uplink scheduler.

12.4.1.1 Downlink Scheduling

The unicast downlink transmission is carried over the shared downlink channel (D-SCH) and the operation takes place at the MAC layer of eNodeB. At each TTI, the eNodeB has to dynamically decide which UE is supposed to transmit, and when and using which frequency resources. The decision depends on different factors including the cell's nominal capacity, QoS parameters (BER, minimum and maximum data rate and delay), backlogged traffic waiting for retransmission, link channel quality relayed to the eNodeB as a CQI, buffers sizes, and the UE's capabilities. More than one UE can be scheduled during one TTI. However, the number of UEs scheduled that can be scheduled during one TTI is limited by the signaling overhead. Allocations are signaled to UEs on the PDCCH, and a UE with enabled downlink reception monitors the PDCCH every TTI.

In addition to the dynamic allocation, LTE standard provides the flexibility to what is called persistent scheduling where the time-frequency resources can be implicitly reused in the consecutive TTIs according to a specific periodicity. Persistent scheduling reduces the overhead scheduling for applications such as VoIP.

Scheduler design is not specified in the standard and is left for vendor implementation. An efficient scheduler, however, should take into account the channel quality of the link from the eNodeB to the UE and the buffer length of the radio bearers. It should also cater to fairness among the UEs based on their service level agreement (SLA), that is, subscription type and priority level. A UE monitors a shared reference signal broadcast to all UEs in the cell by eNodeB to estimate the instantaneous downlink channel quality and signal it in a CQI report. CQI can be about either a single or multiple resource blocks, and can be either periodic or aperiodic. The periodic CQI report is transmitted together with uplink data on the PUSCH or on the PUCCH, while the aperiodic CQI is scheduled by the eNodeB via the PDCCH and transmitted together with uplink data on PUSCH.

12.4.1.2 Uplink Scheduling

The uplink scheduler resides in eNodeB and the UE. Similar to the downlink scheduler, the uplink scheduler at eNodeB is invoked every TTI to decide which UEs will transmit over the uplink shared channel U-SCH, when and using which resources. In addition to assigning the time-frequency resources to the UE, the eNodeB scheduler decides on the modulation and coding scheme that each UE shall use as a consequence of the estimation of the uplink channel quality at the eNodeB.

Fairness, opportunistic (i.e., channel-quality-dependent scheduling), interference coordination and buffer length are performance measures for uplink and downlink scheduling. Considering the buffer size of an uplink radio bearer in scheduling decision to the eNodeB station entails higher overhead and complexity. The UE information about its own radio bearers' buffer sizes is always newer than any signaled information from the UE to the eNodeB. This is one of the reasons for allocating the time-frequency resources per the UE, where in this case the UE will manage the sharing of its uplink resources among its own radio bearers. The Radio Resource Control (RRC) part of the UE MAC layer allocates uplink resources among the radio bearers within the UE. The RRC arbitrates among the radio bearers based on their assigned priorities and an assigned radio bearer parameter called the prioritized bit rate (PBR). RRC first serves the radio bearers in deceasing priority order up to their PBR. Secondly, if there are any residual resources, they are allocated in decreasing priority. In the case that all PBRs are set to zeros the uplink resources are allocated in strict priority order.

To exploit uplink channel quality, the eNodeB requires estimating the uplink channel quality. To achieve this, reference signals, called the channel-sounding reference signals are sent from each UE to the eNodeB. The channel-sounding reference signals are not limited for the frequency resources allocated to the UE, and may span the entire system bandwidth of the cell. Moreover, the channel-sounding reference signals may also be transmitted by UE which does not have any uplink allocated frequency resources.

Both link adaptation at the physical layer and the Hybrid Automatic Repeat Request (HARQ) at the MAC are highly relevant to the scheduling operation. Link adaptation was reviewed in Chapters 2 and 9, while HARQ is explained in the following section.

12.4.2 Hybrid Automatic Repeat Request

LTE provides two mechanisms of error detection and correction through retransmission namely, the HARQ mechanism at L1-MAC and the ARQ at the RLC layer. The ARQ functions less frequently than the HARQ and handles errors not detected by the HARQ process. HARQ is designed to be simple and fast to improve the QoS performance. This improvement is achieved by reducing delay and increasing the system throughput through the fast retransmission. The feedback signal of HARQ is a one bit ACK/NACK and the HARQ can be sent at every TTI.

HARQ is a stop-and-wait ARQ mechanism associated with the unicast transmission on the U-SCH and the D-SCH. HARQ is not employed for broadcast and multicast traffic. The HARQ functionality is terminated at eNodeB – to simplify the architecture, the EPC is isolated from the HARQ procedures. For uplink transmission on the U-SCH, eNodeB decodes the transport block. If successfully decoded, the eNodeB sets the ACK bit in the synchronous feedback signal. The sender identifies the data transmission associated with this ACK signal from the round trip time (RTT) and the timing of the feedback signal. Due to the synchronized feedback, no explicit numbering is required to identify the acknowledged data. Synchronous HARQ applied for uplink transmission is based on scheduling re/transmission of sub-frames at a predefined sequence of time instances. Subframes may be received out of order. Synchronous HARQ transmission is simplified by reducing the control signal overhead and the content of the feedback signal. Additionally, to expedite the HARQ operation, multiple HARQ processes can be concurrently employed for the uplink transmission.

The uplink transmission is triggered by receiving a grant on the Physical Downlink Control Channel (PDCCH). There are two types of grants, grant for new transmission and a grant for retransmission. For a new transmission, the HARQ process uses the uplink grant and the HARQ info (HARQ process ID, New Data Indicator (NDI), Redundancy Version (RV), Transport Block (TB) size) to instruct the PHY layer to transmit the transport block; that is, PDU, set the redundancy version, and stores the Protocol Data Unit (PDU) to be transmitted in the HARQ buffer. The redundancy version is incremental, so in case of retransmission the retransmission are not identical, and a different encoding and data rate can be used.

At eNodeB, HARQ process stores the received transport block in the associated HARQ buffer. If the decoding of this data block is successful the data is handed over to disassembly and demultiplexing, and an ACK is generated and sent. Otherwise, the data is discarded and a NACK is generated and sent if the HARQ

does not support soft combining. If the HARQ does support soft combining, the data is preserved in the HARQ buffer to be used when a new retransmission for the same data is received. The new retransmission is combined with the erroneous PDU in the buffer to generate a single combined PDU. The combined PDU is then decoded. If the decoding fails, a NACK is sent. If the new retransmission is made using an encoding different from the erroneously received PDU, the soft combining used is called the *incremental redundancy,* while if retransmission is identical the soft combining used is the *chase combining.*

When the mobile node receives a grant for retransmission on the PDCCH, the HARQ process uses the retransmission grant retransmission and the HARQ information to generate an adaptive retransmission. It then instructs the PHY layer to retransmit the transport block from the HARQ buffer. If the limit for number of retransmissions is reached, the content of the buffer is flushed and the attempts of retransmitting this PDU are stopped.

At the eNodeB, the retransmitted data is soft combined with the previously received PDU stored in the HARQ buffer. The PDU is then decoded and, if the decoding is successful, the retransmitted PDU is handed for disassembly and demultiplexing and an ACK is generated. Otherwise, a NACK is generated and sent.

The downlink HARQ is also an N-process stop-and-wait ARQ. However, the downlink HARQ is asynchronous, offering flexibility in scheduling retransmissions. This means that the retransmissions are not predefined on specific time instances and can occur at any time instant. This property mandates an explicit HARQ process number to be signaled to associate each retransmission with its HARQ process. Accordingly, retransmissions are scheduled individually, as if they are new transmissions. This enables scheduling retransmissions based on instantaneous radio link conditions. The downlink HARQ also differs in using the incremental redundancy soft combine method with adaptive retransmissions. Adaptive means that the sender can change the transmission attribute as compared to the initial transmission at each retransmission.

ARQ provides higher reliability than HARQ and works less frequently than the HARQ layer, since it is only triggered to correct errors in HARQ operation. HARQ indicates to the ARQ process when the HARQ transmitter reaches maximum retransmissions for a PDU without getting an ACK or when the HARQ detects a transmission failure. The former indication called the NACK1 and the latter is called the NACK2.

When a sender reaches the limit of maximum number of retransmissions the HARQ process at MAC-L1, it sends a NACK1 up to the ARQ process at the RLC layer. In turn, the the ARQ process reworks the transmission block through either segmenting it or using a different encoding. To make sending the NACK1 message possible in uplink retransmission, eNodeB should know the maximum number of retransmissions of each UE to stop generating uplink retransmission grants.

When the HARQ process at the receiver sends NACK and does not receive a retransmission for this NACK in the scheduled TTI, or when the HARQ receives

a new transmission instead, the receiver HARQ process detects that the sender interpreted its NACK erroneously as an ACK. In this case, the receiver HARQ process indicates this to the ARQ process at the RLC layer and the ARQ receiver signals a control message to the sender to resend the erroneously ACK transmission block.

12.5 QoS in LTE-Advanced

Most of the functionalities and specifications related to QoS and radio resource management deployed by LTE are supported by LTE-Advanced to guarantee backward compatibility, which is an essential requirement for the LTE-Advanced standardization. Specifically, QoS performance measures, classification, signaling bandwidth requests and grants are almost similar to LTE. Bandwidth allocation and traffic handling includes some enhancements required to support the new features included in LTE-Advanced to meet or exceed the IMT-Advanced requirements. In this section, we will discuss the major enhancements related to QoS and bandwidth reservation procedures.

12.5.1 Carrier Aggregation

LTE-Advanced provides support for a new feature called Carrier Aggregation, which entails aggregating two or more component carriers that are either contiguous or non-contiguous. The main objective of Carrier Aggregation is to provide larger bandwidth to meet the IMT-Advanced requirements of a spectrum up to 100 MHz.

Carrier Aggregation has an impact on both scheduling and HARQ. For HARQ, it is required in Carrier Aggregation, whether contiguous or non-contiguous, to have one independent HARQ entity per scheduled component carrier. Note that the maximum number of HARQ entities allowed by LTE-Advanced is eight entities for the FDD duplexing. For scheduling, and similar to Release 8, each UE may be simultaneously scheduled over multiple component carriers. However, at most one random access procedure is scheduled per UE in any timeframe. For TDD, it is required that the number of component carriers uplink should be equal to that of the downlink. As in LTE, a single component carrier is still mapped into one transport block.

12.5.2 Coordinated Multipoint Transmission/Reception (CoMP)

CoMP is introduced in LTE-Advanced to mitigate interference and improve the throughput of cell-edge users. CoMP transmission employs dynamic coordination in the scheduling/transmission and/or joint transmission between/from multiple cell sites, while reception employs dynamic coordination in the scheduling and/or joint reception between/at difference cell sites. This enhancement is mainly related to the scheduling function at the eNodeBs participating in CoMP.

There are two types of CoMP:

- Joint processing, and is of two types. In the first, data to a single UE is simultaneously transmitted from eNodeBs participating in CoMP to improve the quality of the received signal at the UE or to actively and dynamically participate in mitigating interference at other UEs. In the second, transmission is performed by one eNodeB in a subframe, where an eNodeB is dynamically selected in each subframe to mitigate interference and improve signal quality at UE.
- Coordinated Scheduling/Beamforming. Here downlink data is transmitted from only the serving eNodeB, but the decisions of when and how to schedule this UE is coordinated with other eNodeBs participating in CoMP.

12.5.3 Relaying in LTE-Advanced

Relaying is currently being studied as an enhancement of LTE towards LTE-Advanced, that is, at the moment it is not part of the standard. The main objective of introducing relaying in LTE-Advanced is to provide extended LTE coverage at low cost. Standardization for relaying is at its early stages and is expected to be finalized by the end of year 2011.

LTE-Advanced relay defines two types of relays, Type-I and Type-II. Type I corresponds to the non-transparent relay in 802.16j, yet differs by being strictly limited to two hops. Type-II corresponds to the transparent relay. In Type-I, the relay node controls its own cell and serves only the purpose of extending the coverage to UEs beyond the eNodeBs effective coverage. A Type I relay node is hence required to transmit the common reference signal and control information to UEs. In Type-II, the UE is within the eNodeB coverage, and is capable of receiving the eNodeB's common reference signal and control information directly. The main objective of Type-II relay node is to increase the overall system capacity by achieving multipath diversity and transmission gains at the UE.

LTE-Advanced relay accommodates different relay transmission schemes to be implemented at the relay node such as:

Amplify and Forward: The simpler transmission mode, and one that is operated at the physical layer. In this mode, the relay station amplifies the signal received from eNodeB (UE) then forwards it to the UE (eNodeB). Type-II relay can employ this transmission mode. While this mode has the advantage of short delay, both the signal and the noise are amplified in the signal relaying process.

Selective Decode and Forward: A relay station employing this transmission mode is capable of limited MAC layer functionalities, specifically channel decoding and cyclic redundancy check (CRC). A relay station decodes the received signal and checks the received message for errors by checking the correctness of the CRC. If the CRC is correct, the relay station performs channel coding then forward then signal to the UE or eNodeB. The selective decode and forward prevents propagating erroneous messages along the path to the UE. It does,

however, incurs longer delays than the amplify and forward due to time required for channel decoding and CRC processing.

Demodulation and Forward: Here, the relay station demodulates the received signal without performing channel decoding or CRC checking. The signal is hence amplified without the noise, and the relay station provides less delay than the selective decode and forward scheme because of the lower processing time. However, it cannot avoid propagating erroneous messages because it does not perform CRC checking.

From the point view of relay architecture LTE-Advanced-relay defines two types of connection between the relay node and eNodeB, namely inband and outband. These connection types are shown in Figure 12.3.

In case of inband connections, the eNodeB-RN link share the same band with direct eNodeB-to-UE. This is applicable for Type II relay. For outband connections, the eNodeB-RN link does not operate in the same band as eNodeB-UE. To enable the inband communications, LTE-Advanced-relay defines resource partitioning procedure with backward compatibility, where network resources are reserved for the eNodeB-RN link and cannot be used by the access link. Figure 12.4 shows the downlink resource partitioning among the relay link and the access link. The resources are time division multiplexed over the same frequency band between the access link and the resource link. A Similar procedure is defined for the uplink communication.

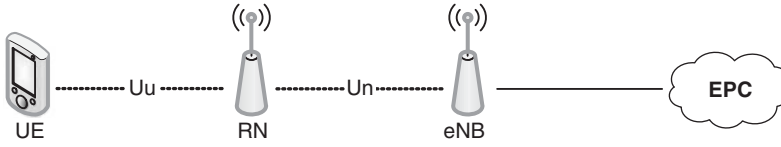


Figure 12.3 LTE-Advanced-relay architecture: redraw to unify the objects used to represent the UE and eNodeB. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

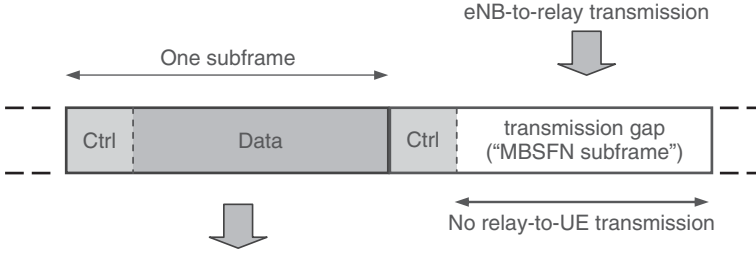


Figure 12.4 Downlink resource partitioning for Inband-connection. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

12.5.3.1 Scheduling

Scheduling in relay mode is noted in TS-36.912 document to be backward compatible and utilizes the procedures and approaches defined in release 8 (LTE). LTE-Advanced-relay defines new downlink physical control channel called the Relay-Physical Downlink Control Channel (R-PDCCH) and two shared channels the Relay-Physical Uplink Shared Channel (R-PUSCH) and Relay-Physical Downlink Shared Channel (R-PDSCH) respectively. Scheduling procedures over these channels are similar to the procedures discussed in Section 14.1.4.1, where semi-persistent and dynamic scheduling are permissible. Scheduling in relay mode can be divided into centralized and distributed scheduling depending on the type of the control provided in the relayed network.

12.5.3.2 Centralized Scheduling

The eNodeB is responsible for scheduling all links of the network, relay links and UE links over the one and two hops distance of the network. The relay node only forwards the received data and signaling from the eNodeB without any scheduling. Global information (channel state information) is hence required at the eNodeB about all network links to engage a centralized scheduling algorithm. To meet the latency requirements of IMT-Advanced and to optimally schedule resources in the network the channel state information of the relay-UE node have to be up-to-date using fast transmission on the backhaul links. Centralized scheduling can be employed in both relaying types.

12.5.3.3 Distributed Scheduling

In LTE-Advanced relaying networks employing distributed scheduling, the scheduler resides at both the eNodeB and the relay node. An eNodeB schedules resources on the eNodeB-RN links and UE directly connected to the eNodeB, and the RN schedule resources on RN-UE links who are two hops distance from the eNodeB. Channel state information of the relay node-UE link need not to be relayed to eNodeB. Consequently, less signaling and overhead is expected in distributed scheduling relay networks. Distributed scheduling can only be employed in Type I relaying networks.

12.5.3.4 HARQ

Two types of HARQ are defined, end-to-end and hop-by-hop. The end-to-end HARQ is simple because the eNodeB has full information about the status of each HARQ transmitted block. HARQ is performed at the eNodeB and the UE, the relay node only relays data and control message between the eNodeB and the UE. The relay node in this case has no contribution in processing. In the event of retransmission/s, the relay node combines the current message with the

previously received messages using the maximal ratio combining then forwards the message. The UE decodes the message and checks for errors using the CRC, and sends back an ACK if the CRC is correct or a NACK if not. The message is forwarded to the eNodeB by the relay node, in case of NACK the eNodeB successively retransmits data corresponding to the same message.

In hop-by-hop HARQ, the relay node not only forwards the data from/to eNodeB/UE, but also contributes in processing. For example, when a relay node receives a message from the eNodeB distant to the UE, the relay node decodes the message, checks the CRC and generates its own feedback (ACK or NACK). When retransmission at the relay node or the UE link occurs, the relay node or the UE, respectively, combines the currently received message with previously received message/s before decoding, decodes the message then forwards the feedback to the eNodeB or relay node respectively. Hop-by-hop HARQ is more efficient than end-to-end HARQ because messages in error are not forwarded and transmission incurs shorter delays. However, decoding and message processing is required by each relay node, which implies that a relay node should support more complex functionality than in the end-to-end HARQ such as buffering and queuing, decoding and CRC checking.

13

Mobility Management

The UE states and state transitions (described in Chapter 11) also dictate the UE and network behavior when it comes to maintaining connectivity. 3GPP have ensured sufficient mechanisms to minimize handover delay and disruption through a simplified RAT and core and management architecture. This enhancement is mainly realized through a signaling hierarchy where the core's involvement in the user's mobility is only made when moving between different management structures (e.g., from one MME to another, or to or from a HeNB) or between RATs. Towards LTE-Advanced, 3GPP has also introduced further handover optimizations that facilitate, for example, soft (zero interruption) handovers.

This chapter is organized as follows. Section 13.1 describes the relationship between the different UE states in 3GPP, and describes transitions between LTE (EUTRAN), UTRAN and GSM. Section 13.2, on the other hand, describes mobility drivers in LTE, identifying the different triggers that would initiate intra- or inter-frequency handovers in LTE, in addition to triggers for inter-RAT handovers. Mobility management for LTE UE is explained in Section 13.3, describing mobility management for both the IDLE and the CONNECTED states. Meanwhile, considerations for Inter-RAT mobility, including procedures for cell reselection and handover, are detailed in Section 13.4. Femtocell or HeNB mobility is reviewed in Section 13.5. Finally, Section 13.6 describes X2 and S1 signaling required for mobility management.

13.1 Overview

Whether idle (RRC_IDLE) or connected (RRC_CONNECTED), a UE's (readiness to) connectivity must be maintained while powered on. Mobility management while IDLE is user-controlled, meaning that it is the UE that seeks the best PLMN and cell to camp on. UE initial selection and ongoing reselection procedures (described in Chapter 11) enable the UE to constantly identify the most appropriate (whether suitable or acceptable) cell or technology

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

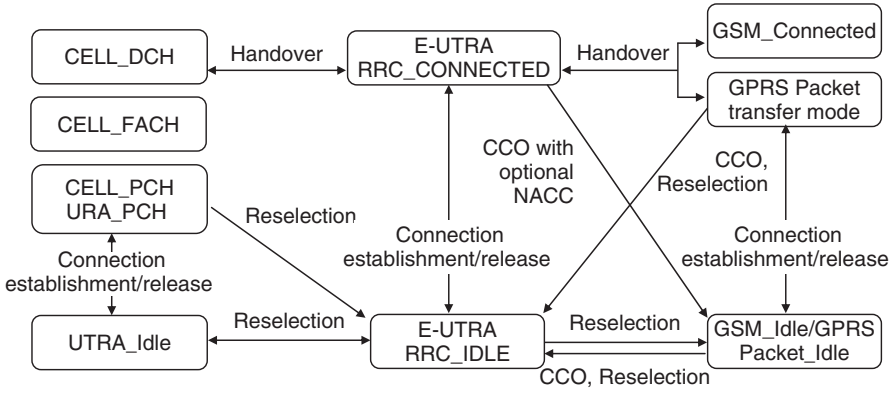


Figure 13.1 States for inter-RAT mobility across 3GPP technologies. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

to camp on. It is the network, however, that takes over this decision when the UE is connected. Once connected, management becomes network controlled but assisted by the UE's measurement reports in selecting best cell, frequency, or RAT that best satisfies the user's requirements and capabilities. Figure 13.1. shows the states for inter-RAT mobility across 3GPP technologies.

As will be described below, 3GPP specifies drivers and limitation for various handoff types that dictate the best action in both states. It also describes the relevant signaling for RRC, X2 (for RAT handovers), and S1 (for core-involved handovers). At times, the network may require additional measurements from the UE for improved decisions. Certain optimizations have also been introduced in Release 9 that enable UE initiated handovers to minimize interruptions for multimedia services.

Resources for this chapter can be sought as follows:

- Drivers and limitations; overall description of RRC services, states and messaging; equivalence between NAS and AS states; overview of S1 and X2 signaling; descriptions of HeNB relevant aspects; description of certain handover optimizations can all be found in 36.300.
- Descriptions for RRC processes in both IDLE and CONNECTED states and details of measurement requests and procedures are described in 36.331.
- Network requirements for supporting radio resource management are described in 36.133.
- S1AP and X2AP are respectively described in 36.413 and 36.423.

13.2 Drivers and Limitations for Mobility Control

Table 13.1 the driver and limitations for handover decisions, in addition to their applicability for different mobility scenarios. Drivers form the operational basis

Table 13.1 LTE handover drivers and limitations, and applicability to mobility scenarios. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited

	#	Drivers/limitations	Intra-frequency	Inter-frequency	Inter-RAT
Drivers	1	Best radio condition	X	X	X
	2	Camp load balancing		X	X
	3	Traffic load balancing		X	X
	4	UE capability		X	X
	5	Hierarchical cell structures		X	X
	6	Network sharing		X	X
	7	Private networks/home cells		X	X
	8	Subscription based mobility control		X	X
	9	Service based mobility control		X	X
	10	MBMS		X	X
Limitations	11	UE battery saving	X	X	X
	12	Network signalling/ processing load	X	X	X
	13	U-plane interruption and data loss	X	X	X
	14	OAM complexity	X	X	X

upon which a handover decisions should be made, while limitations indicate to aspects that may constrain the selection of the handover type or scenario. For LTE, 3GPP distinguishes between three types of mobility scenarios. The first, intra-frequency, is the fundamental handover scenario and is strictly driven by "best radio condition" driver – no other driver can result in an intra-frequency handover as it will definitely result in a degrading performance. Inter-frequency, on the other hand, becomes possible when an operator has simultaneous access to multiple carriers or bands for LTE. This accessibility can be either fixed or temporary. The resulting flexibility allows for different decisions in resource control and architecture, that is, dedicate certain bands for certain services, or establish a network hierarchy. Similarly, the various drivers and limitations can be applied if the operator (or user) has access to multiple RATs.

As aforementioned, the "best radio condition" driver supersedes any other driver for handover decisions. This is achieved through cell reselection, and the standard provides – where possible – sufficient measures in terms of signaling, operations and measurements to maintain this driver's objective. For handovers to different frequencies or RATs, the UE should be provided sufficient opportunity to verify handover viability.

The remaining drivers address various needs for handovers that may arise for network operation. For example, load balancing idle or connected UEs between

different bands or RATs may be possible, depending on UE capabilities. The UE capability driver is a generic driver, allowing for supporting opportune overlay and UE capabilities combinations. The driver "private networks/home cells" enables users with accessible HeNBs, that is, either the user is in the CSG or the HeNB is open, to handover when possible. Care, however, should be taken with such driver to avoid unnecessary handoffs.

Handover limitations including UE's battery, signaling and processing overload, U-plane interruption and loss possibilities, and OAM complexity. These limitations ensure that handover decisions do not result in the UE unjustifiably expending battery energy, nor result in excessive signaling or processing on part of the network or the UE. Also, a handover decisions that risks serious interruption to user service delivery should be avoided. Finally, all handover decisions should be made so that they do not require excessive efforts in operation, administration, or management of network resources.

Beyond describing the drivers and limitations, the standard also describes required features to support the various handover drivers and scenarios in the IDLE and ACTIVE modes, and in transitioning between. The required features are described in full for handovers made across the different 3GPP technologies (LTE, UTRAN, and GERAN.) For example, in transitioning between IDLE to ACTIVE in LTE and between LTE and UTRAN, the required features include inter-frequency/RAT measurements and measurement reporting upon RRC establishment. Supporting the "traffic load balancing" in the ACTIVE mode requires the viability of a network controlled inter-frequency/RAT handover, and exchange of load information across the different frequencies/RATs. These and other descriptions are made detailed in Annex E in 36.300.

13.3 Mobility Management and UE States

13.3.1 IDLE State Mobility Management

The initial entry procedure, described in Chapter 11, details PLMN selection and the cell selection steps required. Based on interactions between the NAS and the AS, the UE is able to identify the PLMN, and search the E-UTRA's frequency band to identify – for each carrier frequency – the strongest cell. The objective is to first a suitable cell for the UE to camp. If no suitable cells are available, the UE tries to find an acceptable cell – one through which it is able to initiate emergency calls, and receive ETWS and CMAS messages. Coming from RRC-CONNECTED, the UE would either camp on the cell through which it was connected, or the cell indicated by the RRC in the state transition.

Once in the RRC-IDLE, the UE continues reselection. UE makes measurements of the serving and neighboring cells to enable to the reselection process. If the UE is either "camped normally" or "camped on any cell", it detects, synchronizes and monitors intra-frequency, inter-frequency and inter-RAT cells indicated by the serving cell. The network need not support this measurement activity, that is,

through explicitly providing a list of neighboring cells and their frequencies and bandwidth information.

Intra-frequency reselection is based on ranking of cells in terms of the average signal quality. Inter-frequency reselection is based on frequency priorities set by the PLMN. The serving cell can provide a Neighboring Cell List (NCL) for intra- and inter-frequency neighboring cells. It can also provide black lists to prevent the UE from reselecting to specific intra- and inter-frequency neighboring cells. If any cell reselection parameters are provided in a cell, they become applicable to all UEs within that cell.

13.3.2 CONNECTED State Mobility Management

When a UE is at the RRC_CONNECTED state (i.e., ECM-CONNECTED with radio bearers), the network handles the UE's handover decisions, including evaluation of eNB measurements and UE measurements limitations, communication with target cell or network, informing the U of new radio resources and releasing unused resources. The network also oversees mechanisms for context transfer and updating node relations on C-plane and U-plane.

As in cell reselection, the UE makes measurements of attributes of the serving cells and neighboring cells and networks. The eNB need not indicate to the UE neighboring cells. It need, however, to indicate the carrier frequencies of the neighboring of inter-frequency neighboring cells. The eNB can provide an NCL or black lists of neighboring cells.

Whether or not an UE requires a measurement gap depends on the carrier frequency of neighboring cells. To elaborate, if both the serving and target cells have the same carrier frequency, the UE does not require a measurement gap. This is regardless of whether or not the bandwidths of the two cells completely overlap. If the carrier frequencies of the two cells are different, then the UE requires a measurement gap. The different instances are shown in Figure 13.2.

An intra EUTRAN handover while the UE is RRC_CONNECTED does not require the involved of the EPC. The handover command sent by the serving eNB to the UE comes from the target eNB. The serving eNB, in preparation for the handover, would transfer the relevant necessary information, for example, UE's context. Both the eNB and the UE maintain some context in case the handover fails. The UE's access to the target cell is made using RACH in a contention-free procedure. This means that the UE requires a dedicated RACH preamble. This preamble is used until the handover procedure is finish.

As will be explained later on in the chapter, there is a hierarchy of handover types. The most basic handover type and one that does not require the involvement of the EPC is when the UE is handed over between MME without changing its serving gateway. A successful instance of such a handover is shown in Figure 13.3.

In the figure, once the Source eNB receives the UE's measurement reports and decides that a handover would be appropriate, it communicates a Handover Request to the target Enb. Upon receiving the request, the target eNB performs

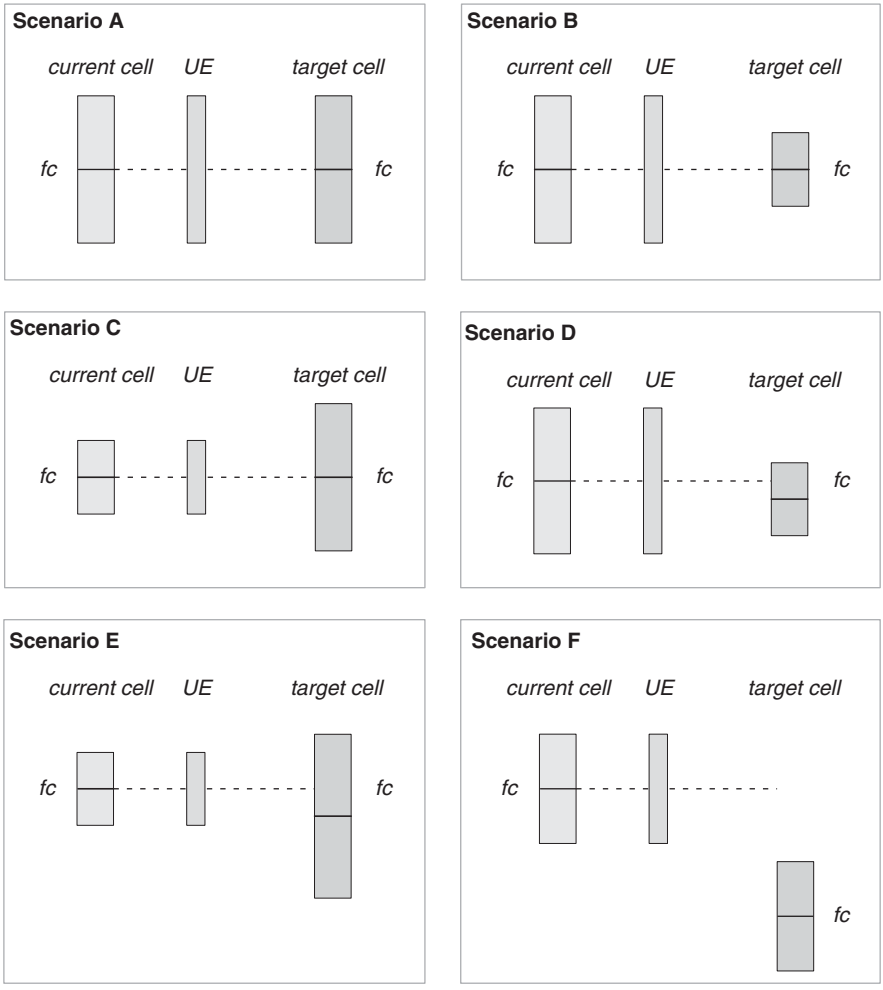


Figure 13.2 Inter and intra-frequency measurement scenarios. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

admission control to judge whether it can handle the UE's requirements. If sufficient resources are available, the target eNB acknowledges the handover requests, including information such as random access preamble, downlink allocation, etc, which are transparently relayed by the source eNB. The Source eNB also issues an RRC connection reconfiguration message. Upon receiving the reconfiguration message, the UE begins detachment from the source eNB and begins synchronization with the target Enb. At this instance, the source eNB begins forwarding the UE's data and transfer the UE's context to the target Enb.

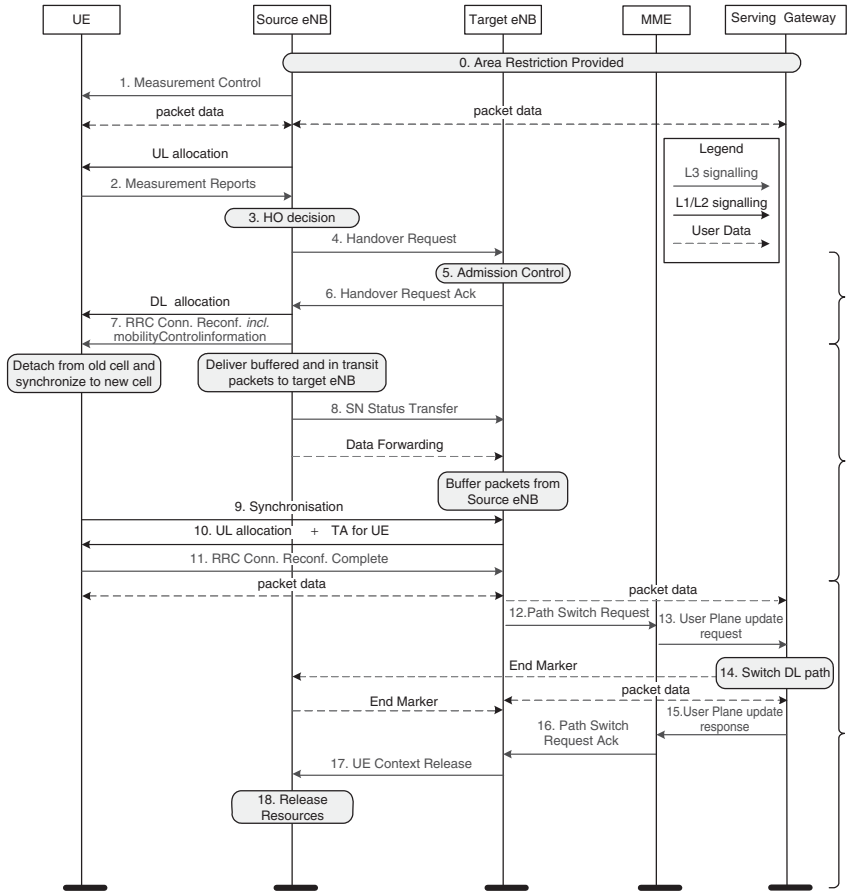


Figure 13.3 Intra-MME/Serving Gateway HO. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

The above procedure is completely performed at the RAT level. Once completed, the target eNB communicates with the MME to simply switch the UE's communication path. The MME in turn updates the S-GW as to the UE's user plane new status. The S-GW then issues a switch DL path command and immediately begins newly arriving data to the target eNB. Signaling is then exchanged such that the target eNB's path switch request is acknowledge, and the source eNB releases the UE's resources.

13.4 Considerations for Inter RAT Mobility

The standards LTE and LTE-Advanced provided detailed descriptions for inter-RAT mobility, both to and from other 3GPP and non-3GPP technologies

(respectively, GERAN/UTRAN and cdma2000/HRPD). A brief overview of cell reselection and handover procedures for 3GPP is provided below.

13.4.1 Cell Reselection

A UE can only search and measure for neighboring GERAN cells if their details are provided in the serving cell's NCL. For UTRAN, however, the serving cell can provide a list of carrier frequencies. A UE, in its continuing search for a better cell (frequency, technology) to camp on, selects the RAT with the highest priority. Priorities are set by registered PLMN (i.e., a PLMN that the UE has successfully registered on), and are valid only within that PRLMN.

If the UE is camped on another RAT, the UE need to acquire the carrier frequencies of the neighboring EUTRAN cells in order to be able to search and measure. There is no need to indicate cell-specific reselection parameters since such parameters are common to all neighboring cells on an E-UTRA frequency.

13.4.2 Handover

A basic principle in LTE and LTE-Advanced is that handovers to GERAN and UTRAN are minimized. Such handovers are controlled through the source access system, which decides whether or not to initiate handover and, when a handover is initiated, provides sufficient information to the target system. They are described as being "backward handovers", meaning that the target system must acknowledge that sufficient resources are available and ready for the incoming UE prior to handover execution. To facilitate these backward handovers, interfaces are defined between the corresponding MME/S-GW and the 2G/3G SGSN. Moreover, it is responsibility of the target access system to provide specific guidance us on how to make the radio access there. The relevant information are sent by the target system are transparently relayed by the source system. At the same time, the handover should not require any UE to core (CN) communication to redirect flows to the target system.

13.5 CSG and Hybrid HeNB Cells

In presence of allowed cells, a UE performs reselection for intra-frequency mobility as if with regular cells. An allowed cell can either be on in the UE's CSG whitelist, or a hybrid cell. In ranking and reselection, the UE may ignore all CSG cells that are known to be not allowed. In case of inter-frequency mobility, the UE prioritizes CSG cells that are the UE's whitelist, irrespective the general network priorities for frequencies. As for inter-RAT, LTE supports both inbound and outbound mobility.

When IDLE, and in instances where different CSG HeNB use different mixed carrier, all CSG cells broadcast the PCI values reserved by the network for use by CSG cells. This broadcast is optional. It is also optional that non-CSG cells

on the mixed carrier can send this information in system information. A UE checks the suitability of CSG cells based on whether they're the UE's whitelist. Manual selection of CSG cell is supported.

When CONNECTED, the UE performs measurement and mobility procedures as set by the network. A UE is not required to support manual CSG selection when connected.

The above descriptions apply for mobility inbound to a CSG or hybrid cell. For outbound mobility, a UE performs normal IDLE mode reselection and CONNECTED mode handover procedures.

The general procedure for a handover from eNB/HeNB to a HeNB's CSG or hybrid cell is shown in Figure 13.4. When the Source EUTRAN cell (eNB or HeNB) decides that a handover is required, a Handover Required message is relayed to the MME carrying the target cell's global identity and the CSG ID. If the target is a hybrid cell, the Cell Access Mode is also included. The MME performs UE access control to the CSG cell based on the CSG ID. If the access control procedure fails, the MME ends the handover procedure with a Handover Preparation Failure message. If successful, the MME relays a Handover Request to the HeNB's gateway, which in turn relays the request to the target HeNB. Once the target HeNB validates the CSG ID, it acknowledges the Handover Request to the MME through its gateway. In turn, the MME sends the Handover Command message to the source EUTRAN cell.

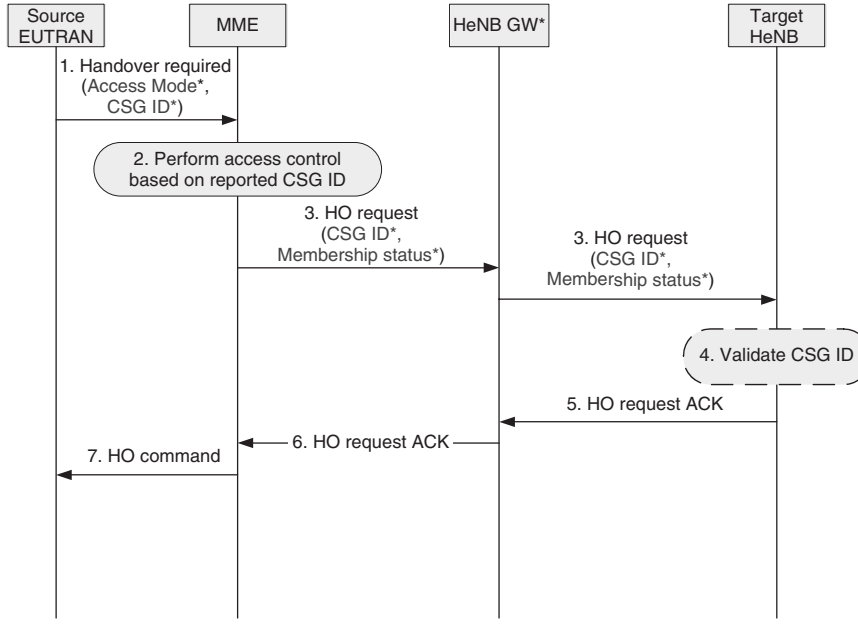


Figure 13.4 Intra eNB/HeNB to HeNB HO. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

13.6 Mobility Management Signaling

Intra-RAT (inter-eNB) handovers are handled by the X2 interface unless any of the following conditions are true:

- No X2 interface between the source and target eNBs.
- The source eNB has been configured to initiate handover to the particular target eNB via the SI interface in order to enable the change of en an EPC node (MME and/or Serving GW).
- The source eNB has attempts to start an inter-eNB handover via X2 but receives a negative reply from the target eNB with a specific cause value.
- The serving PLMN changes during handover.

The S1 interface handles inter-RAT handovers. It also handles handovers from and to HeNBs.

In what follows, the relevant signaling for the X2 and S1 interfaces are described.

13.6.1 X2 Mobility Management

Elementary procedures for mobility management in the X2 interface include: Handover Preparation; SN Status Transfer; UE Context Release; and Handover Cancel.

Handover Preparation involves the source eNB sending a Handover Request to the target eNB. When the target eNB receives the request, it performs an admission control request based on the resources indicated to be required for the UE's radio bearers. If sufficient resources are available, the target eNB acknowledges the request with a Handover Request Acknowledge. If not, the target eNB sends a Handover Preparation Failure message to the source eNB. This terminates that Handover Preparation procedure. If beyond a certain time the source eNB does not receive an indication of either an acknowledge or a failure, it sends a handover Cancel and indicates the cause as expired timer. If a source eNB sends cancels a handover requests, it disregards further messages from the target eNB.

The SN Status Transfer message is used to transfer the uplink PDCP SN and HFN receiver status and the downlink SN and HFN transmitter status from the source to the target eNB during an X2 handover for each respective E-RAB for which the PDCP SN and HFN status preservation applies. The status message includes information on missing and received uplink SDUs, and for downlink flows for which it has received reports on missing and received SDUs.

The release of UE context, by which the source eNB release all information relative to a UE, is made by the target eNB sending a UE Context Release to the source eNB. It is also an indication of handover success. It is possible, however, that the source eNB would retain the UE's context after receiving the context release message in case of failure.

The Handover Cancel message is sent by the source eNB to the target eNB to cancel an ongoing handover preparation or an already prepared handover. The cancel message carries reason for cancellation. Upon receiving the message, the target eNB releases any resources or context relevant to the UE.

There are two possible X2-based handovers: without or with Serving GW relocation. Figure 13.5 shows the procedure for an X2-based handover where the MME decides that the Serving GW will not be changed.

After Handoff preparation and execution, which were described above, the Handover Completion stage entails the target eBN sending a Path Switch Request. A request and response for modifying the UE's radio bearers are then processed by the Serving GW and, in turn, the PDN GW. Once complete, the Serving GW responds to the path switch request by redirecting the UE's downlink data, sending an end marker to the source eNB. In turn, the source eNB sends another end marker to the target eNB and the MME acknowledges the target eNB's path switch request. Finally, the target eNB sends a Context Release message to the source eNB.

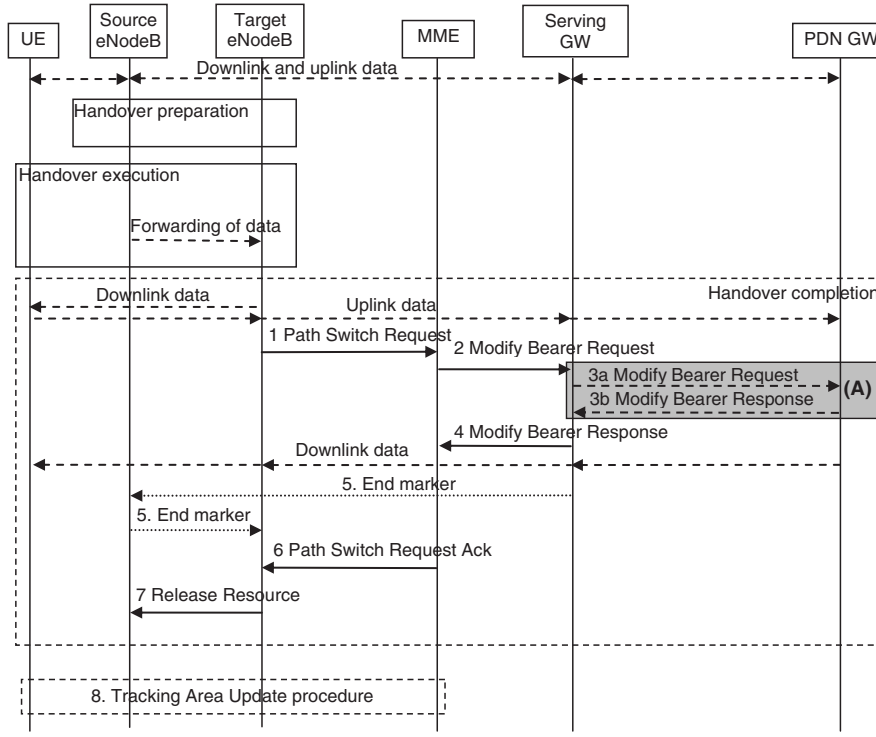


Figure 13.5 X2-based handover without relocating serving GW. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

When the Serving GW is relocated during an X2-based handover, additional messages are exchanged between the MME and both the source and the target Serving GWs. First, the MME establishes a session with the target Serving GW, which in turn exchanges the bearer modification messages with the PDN GW. The MME then acknowledges the target eNB's path switch request, and deletes

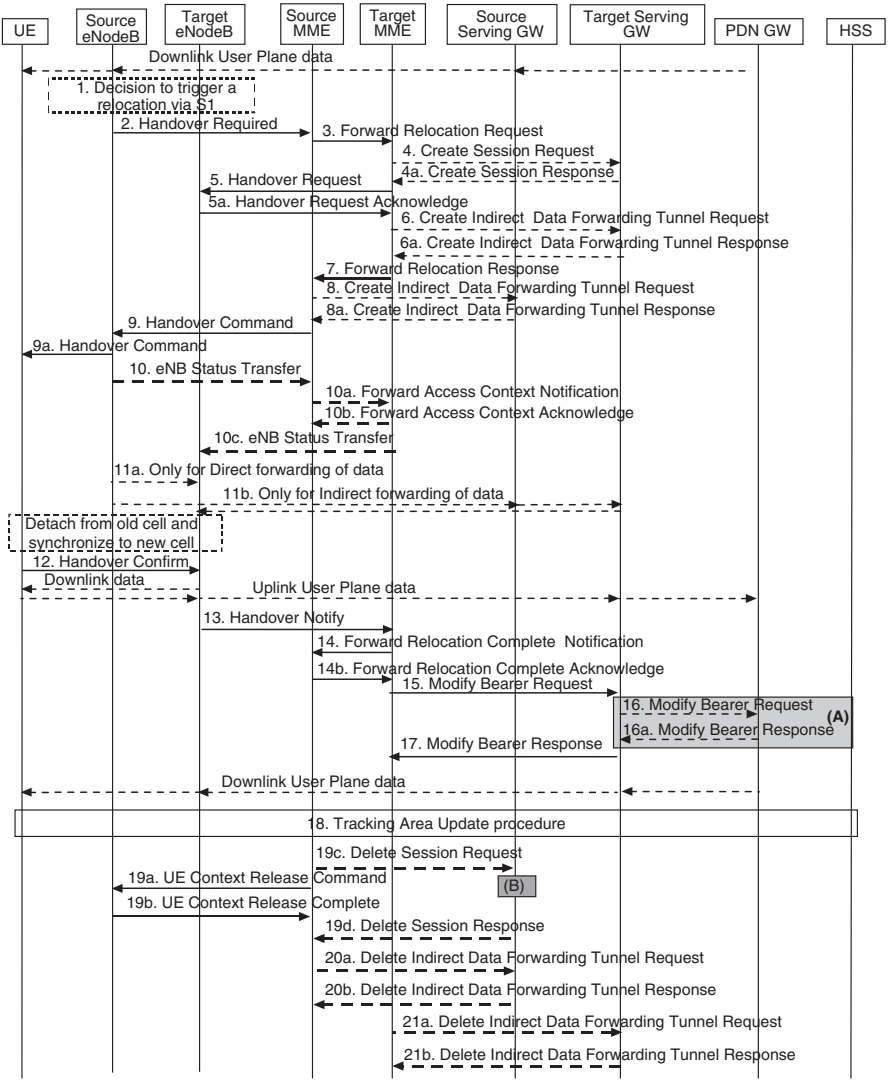


Figure 13.6 A General S1-based handover. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

the session with the source serving GW. Meanwhile, the target eNB would send the source eNB a Context Release message.

13.6.2 S1 Mobility Management

S1 mobility management oversees handovers that cannot be initiated by the X2 interface. They also handle HeNB handovers inter-RAT handovers.

There are several handover related signaling in the S1 interface. Handover Required, Handover Request, Handover Notify, Path Switch Request, Handover Cancel, eNB Status Transfer, and MME Status Transfer.

A Handover Required is sent by the source eNB to the MME to initiate an S1 handover preparation phase. If the MME judges that the handover can be realized, it sends a Handover Command. Otherwise, it sends a Handover Preparation Failure. When processing the Required message, the MME considers the source and target eNBs (their RATs, frequencies, resources) and the UE capabilities. Consideration for access control for HeNB are also made at the MME. An indication of a successful handoff is sent by the target eNB to the MME using a Handover Notify message.

An MME performs a handover resource allocation by sending a Handover Request message to a target eNB. Upon receiving the request message, the target eNB makes the appropriate resource allocation, context preparation and relevant security authentications for the UE. If the target eNB is unable to admit the UE, it responds with a Handover Failure message.

The Path Switch Request message is exchanged between the eNB and the MME for the MME to process the path switch with the core. Both status transfer exchanges (the eNB and MME Status Transfer) indicate the PDCP SN and HFN status in the uplink and the downlink. Finally, a handover initiated by a source eNB can be cancelled at any time for an ongoing or a prepared for handover using a Handover Cancel message. Upon receiving the cancel message, the MME would acknowledge the cancellation and initiate relevant release and removal procedures.

A general S1-based handover is shown in Figure 13.6.

14

Security

The separation of user and control planes and the access and non-access stratums in LTE/SAE result in an implicit security requirement. 3GPP describe an extensive two layer security architecture that also utilizes IETF security solutions for its IP core. The security architecture is maintained in LTE-Advanced, with some enhancements concerning more capable encryption and integrity algorithms being utilized.

This chapter is organized as follows. Section 14.1 offers the rationale behind the design of 3GPP's security architecture for both LTE and LTE-Advanced. Section 14.2 describes the security architecture, including security features defined by the standard for network access and domain, user domain, application domain and the visibility and configurability of security. The key hierarchy in EPS is relevant to LTE, and is hence explained in Section 14.3, while the relationship between the UE states and state transitions and securities are outlined in Section 14.4. Finally, Section 14.5 describes the security procedures that take place between the UE and elements at the network core.

14.1 Design Rationale

In Chapter 9, it was discussed how it was essential for LTE/SAE to separate the communication between the core and the UE to the AS and the NAS. This separation was also applied in terms of security whereby the security of the AS (i.e., RRC security in eNB) was separated from the security of the NAS signaling. Two other relevant decisions were made in the design of the security architecture for LTE. The first is that the user plane security terminates above the eNB; the second, that the radio link and the core network must have

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

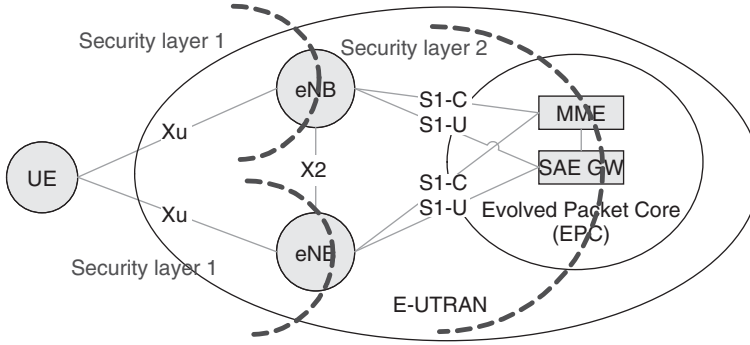


Figure 14.1 First and second security layers in LTE. Reproduced by permission of © 2009 3GPP. Further use is strictly prohibited.

cryptographically separate keys. These requirements result in LTE having two layers of protection, differentiating the E-UTRAN from the UTRAN which only a one layer perimeter security. These two layers are shown in Figure 14.1, where the first provides the RRC security and the User plane protection, while the second layer provides the NAS signaling security.

The immediate advantage of this rationale is that a compromise at the first layer (i.e., if an eNB or an HeNB is compromised), it would be hard to compromise the security of the eNB/HeNBs (i.e., other elements in the first layer) or the core (i.e., layer 2). This means that placing eNBs in vulnerable locations is more practically accessible in LTE.

Resources for this chapter can be sought as follows.

- The rationale for LTE/SAE security architecture can be found in 33.821.
- Overview of security architecture, in addition to details on access and core security procedures for 3GPP accesses can be found in 33.401.
- Security architecture and procedures for non-3GPP accesses are described in 33.402.
- Descriptions of user-side USIM, application and visibility/configurability can be found in 33.102.
- Descriptions of RRC security signaling and activation are described in 36.331.

14.2 LTE Security Architecture

Figure 14.2 gives an overview of the complete security architecture for LTE. The stratums identified, each addressing a sufficiently isolated category of security threats, are the application, home, serving and transport stratum.

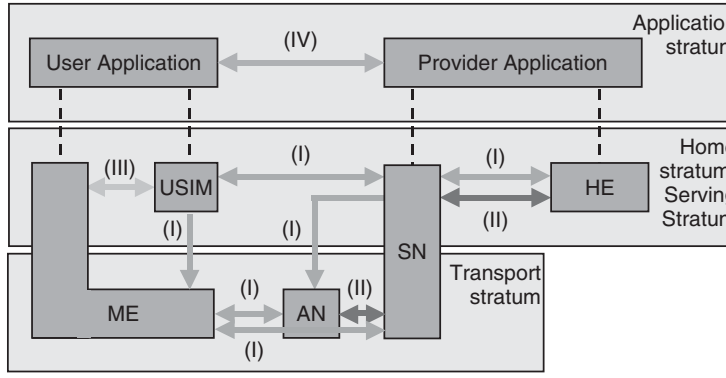


Figure 14.2 Overview of LTE Security Architecture. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

As can be noted in the figure, there are five sets of security features the 3GPP define:

(I) *Network access security*: The set of security features that provide users with secure access to services, and which in particular protect against attacks on the (radio) access link.
(II) *Network domain security*: The set of security features that enable nodes to securely exchange signaling data, user data (between the Access Network (AN) and the Serving Network (SN), and within the AN), and protect against attacks on the wireline network.
(III) *User domain security*: The set of security features that secure access to mobile stations.
(IV) *Application domain security*: The set of security features that enable applications in the user and in the provider domain to securely exchange messages.
(V) *Visibility and configurability of security*: The set of features that enables the user to inform himself whether a security feature is in operation or not and whether the use and provision of services should depend on the security feature.

In what follows, we elaborate on some of these feature sets.

Network access security entails specific feature such as user identity confidentiality, entity authentication, general confidentiality of certain agreement and data exchanges, and data integrity. Identity confidentiality is normally achieved by assigning short-lived temporary identities to ensure confidentiality of both user identity and location, and user untraceability. Meanwhile, entity authentication applies to both user and network authentication. Realizing entity authentication is

Table 14.1 Security Termination Points. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited

	Ciphering	Integrity Protection
NAS Signalling	Required and terminated in MME	Required and terminated in MME
U-Plane Data	Required and terminated in eNB	Not Required (NOTE 1)
RRC Signalling (AS)	Required and terminated in eNB	Required and terminated in eNB
MAC Signalling (AS)	Not required	Not required

NOTE 1: Integrity protection for U-Plane is not required and thus it is not supported between UE and Serving Gateway or for the transport of user plane data between eNB and Serving Gateway on S1 interface.

made possible through authentication at each connection set up between the network and the user. General confidentiality applies to cipher algorithm and key agreements, and user and signaling data. Finally, integrity algorithm and key agreements, in addition to data integrity and origin authentication of signaling data are all properties achieved various mechanisms.

Ciphering may be provided to RRC-signaling to prevent UE tracking on over-the-air RRC exchanges, for example, for measurements or handover. NAS signaling may also be confidentiality protected. Confidentiality of user plane exchanges should be made at the PDCP layer. This measure, however, is optional. Meanwhile, integrity shall be provided (i.e., is mandatory) for both NAS and RRC-signaling. These measures will be described below. Table 14.1 shows the termination points for the NAS signaling, U-plane, and the AS (RRC and MAC signaling)

Network domain security refers to general IP-relevant security measures that apply various IETF syndicated measures. These measures are detailed in further details in 33.210 and 33.10 (respectively describing security aspects for IP network layer and the network domain authentication framework).

User domain security involves user-to-USIM authentication, and authorization of the USIM-Terminal link. These are basic security measures to authenticate any user or terminal. Meanwhile, application security is enabled by the security features provided for the USIM Application Toolkit which enables authentication applications residing the USIM.

Note that a similar architecture is assumed when dealing with non-3GPP accesses, where the access and serving networks would be a non-3GPP access network.

14.3 EPS Key Hierarchy

Two requirements bound the EPS key hierarchy and derivation. The first is that the EPC and E-UTRAN shall allow for use of encryption and integrity protection

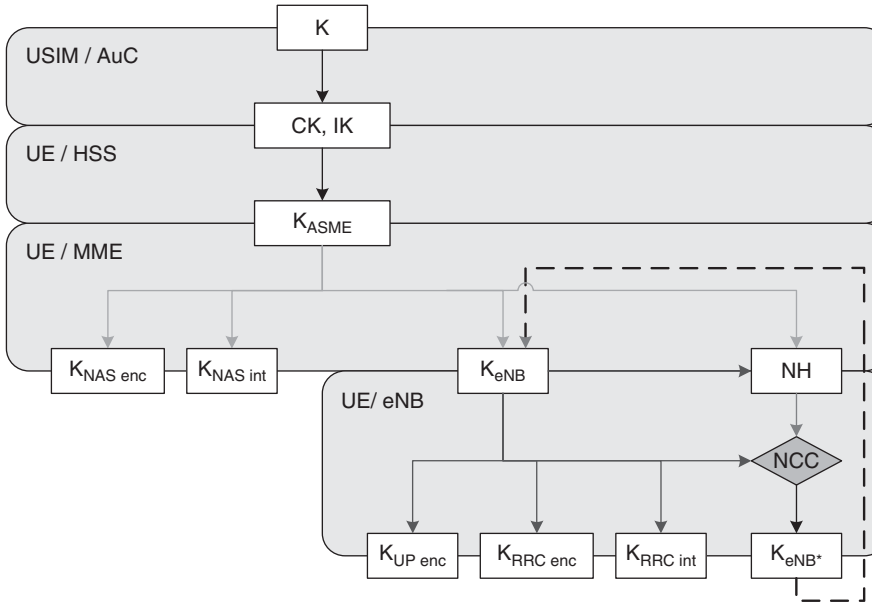


Figure 14.3 EPS Key Hierarchy and Derivation. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

algorithms for AS and NAS protection having keys of length 128 and for future use the network interfaces shall be prepared to support 256 bit keys. The second is that keys for the user plane, NAS and AS protection shall be dependent on the algorithm with which they are used.

The hierarchy, shown in Figure 14.3, includes the following keys: K_{eNB}, K_{NASint}, K_{NASenc}, K_{UPenc}, K_{RRCint} and K_{RRCenc}. A brief description of the different keys and how they are derived is provided below.

- **K_{eNB}** is a key derived by UE and MME from K_{ASME}. K_{eNB} may also be derived by the target eNB from NH at handover. K_{eNB} shall be used for the derivation of K_{RRCint}, K_{RRCenc} and K_{UPenc}, and for the derivation of K_{eNB*} upon handover.

Keys for NAS traffic:

- **K_{NASint}** is a key, which shall only be used for the protection of NAS traffic with a particular integrity algorithm. This key is derived by UE and MME from K_{ASME}, as well as an identifier for the integrity algorithm.
- **K_{NASenc}** is a key, which shall only be used for the protection of NAS traffic with a particular encryption algorithm. This key is derived by UE and MME from K_{ASME}, as well as an identifier for the encryption algorithm.

Keys for UP traffic:

- **K_{UPenc}** is a key, which shall only be used for the protection of UP traffic with a particular encryption algorithm. This key is derived by UE and eNB from K_{eNB}, as well as an identifier for the encryption algorithm.

Keys for RRC traffic:

- **K_{RRCint}** is a key, which shall only be used for the protection of RRC traffic with a particular integrity algorithm. K_{RRCint} is derived by UE and eNB from K_{eNB}, as well as an identifier for the integrity algorithm.
- **K_{RRCenc}** is a key, which shall only be used for the protection of RRC traffic with a particular encryption algorithm. K_{RRCenc} is derived by UE and eNB from K_{eNB} as well as an identifier for the encryption algorithm.

Intermediate Keys and Values:

- Next Hop (**NH**) is used by UE and eNB in the derivation of K_{eNB*} for the provision of "forward security". NH is derived by UE and MME from K_{ASME} and K_{eNB} when the security context is established, or from K_{ASME} and previous NH, otherwise.
- Next Hop Chaining Count (**NCC**) is a counter related to NH (i.e., the amount of Key chaining that has been performed) which allow the UE to be synchronized with the eNB and to determine whether the next **K_{eNB*}** needs to be based on the current K_{eNB} or a fresh NH.

Forward security: In the context of K_{eNB} key derivation, forward security refers to the property that, for an eNB with knowledge of a $K_{eNB,}$ shared with a UE, it shall be computationally infeasible to predict any future $K_{eNB,}$ that will be used between the same UE and another eNB. More specifically, n hop forward security refers to the property that an eNB is unable to compute keys that will be used between a UE and another eNB to which the UE is connected after n or more handovers (n = 1 or 2).

14.4 State Transitions and Mobility

A UE transitioning between RRC_IDLE to RRC_CONNECTED must have its RRC and UP protection keys generation while NAS and higher layer protection keys are assumed to be already available in the MME. Higher layer keys may have been established in the MME as a result of an AKA run, or as a result of transfer from another MME during handover or idle mode mobility.

When transitioning between RRC_CONNECTED to RRC_IDLE, eNBs delete all the keys they store such that the state for IDLE mode has be maintained only at in the MME. The eNB will also not be storing any state information about the corresponding UE. Specifically, both the eNB and the UE will delete NH,

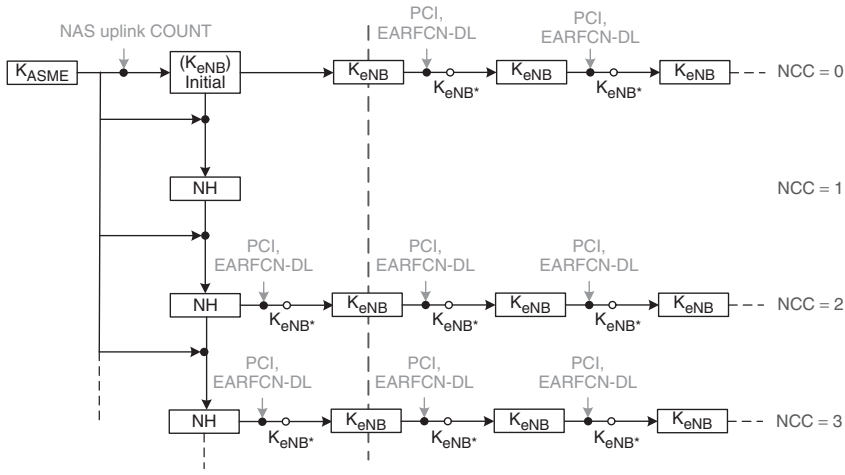


Figure 14.4 Key Handling during Handover. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

K_{eNB}, K_{RRCenc}, K_{RRCint}, K_{UPenc} and related NCC, but the MME and the UE will maintain the K_{ASME}, K_{NASint} and K_{NASenc}.

During mobility, the key hierarchy does not allow explicit RRC and UP key updates, but RRC and UP keys are derived based on algorithm identifiers and K_{eNB} which results with new RRC and UP keys at every handover. Figure 14.4 shows the model for key handling during handover. The handling proceeds as follows.

Whenever an initial AS security context needs to be established between UE and eNB, MME and the UE shall derivate a K_{eNB} a NH, both of which are derived from the K_{ASME}. The UE and the eNB use the K_{eNB} to secure the communication between each other. On handovers, the basis for the K_{eNB} that will be used between the UE and the target eNB, called K_{eNB}*, is derived from either the currently active K_{eNB} or from the NH parameter. The former derivation is called a horizontal key derivation, while the latter is called a vertical key derivation. On handovers with vertical key derivation the NH is further bound to the target PCI and its frequency (EARFCN-DL) before it is taken into use as the K_{eNB} in the target Enb. On handovers with horizontal key derivation the currently active K_{eNB} is further bound to the target PCT and its frequency before it is taken into use as the K_{eNB} in the target eNB.

14.5 Procedures between UE and EPC Elements

14.5.1 EPS Authentication and Key Agreement (AKA)

The EPS AKA produces keying material forming a basis for the user plane, RRC and the NAS ciphering keys as well as RRC and NAS integrity protection keys.

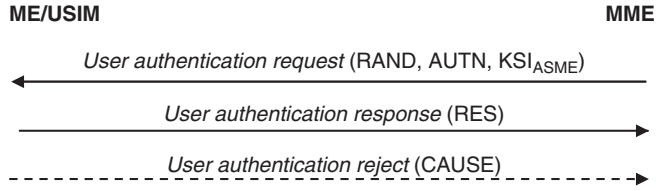


Figure 14.5 EPS Authentication and Key Agreement. Reproduced by permission of © 2010 3GPP. Further use is strictly prohibited.

The MME sends to the USIM a random challenge, an authentication token, in addition to the K_{ASME}. The K_{ASME} key is a base key, from which NAS keys and K_{eNB} keys and H are derived. The K_{ASME} is never transported to an entity outside of the EPC, but K_{Enb} and NH are transported to the eNB from the EPC when the UE transitions to ECM-CONNECTED. From the K_{eNB}, the eNB and UE can derived the UP and RRC Keys.

When the USIM receives the authentication request, as shown in Figure 14.5, it verifies the freshness of the authentication vector and, if acceptable, computes a response. If the verification fails, the ME responds an authentication reject message indicating cause.

14.5.2 Distribution of Authentication Data from HSS to Serving Network

This procedure enables the HSS in the UE's home environment to provide one or authentication vector to the serving network's MME to perform user authentication. The standard recommends that only one EPS authentication vector is fetched due to capability of an elaborate key hierarchy (see below). The authentication data request shall include the IMSI, serving networking identity and the network type. When the HE receives the request, can use either pre-computed or vectors or compute vectors on demand.

14.5.3 User Identification by a Permanent Identity

The user identification mechanism is invoked by the serving network whenever the user cannot be identified by means of a temporary identity, especially when the serving network cannot retry the IMSI based on the GUITI by which the user identifies itself on the radio path.

Part Three

Comparison

15

A Requirements Comparison

This chapter provides a comparative study between LTE and WiMAX. Surfing the Internet looking for details about the two technologies unveils a fierce competition. However, we can safely say that this competition is confined to sharing the wireless broadband market rather than a technological competition. The reason behind this, as was established in the previous chapters, is the many similarities between the PHY and MAC functionalities used in the two. In fact, they adopted the same technologies and only differed in implementation. As a result, it is not expected that one of them will eliminate the other, but they will rather smoothly integrate after the market wars settles down. Therefore, it is logical to provide a comparative instead of a competitive study of the two.

This chapter is organized as follows. Section 15.1 contrasts the evolution towards the two IMT-Advanced technologies. Section 15.2 is dedicated to comparing the spectral efficiency performance, and comments on the implementation and adoption of both OFDMA and MIMO. Spectrum flexibility and the application of carrier aggregation techniques are discussed in Section 15.3. Finally, Section 15.4 compares network architectures.

15.1 Evolution of the IMT-Advanced Standards

Despite the many similarities between LTE and WiMAX, their origins are radically different. While LTE is the inheritor of a voice-service based technology, namely the voice circuit switching, WiMAX originated from a data-service based technology, the computer networks. In other words, WiMAX was born as an all IP technology while LTE went through a slow and long development stages to become an all IP network. The evaluation of both standards is shown in Figure 15.1.

LTE predecessors can be traced back to the circuit-switched analog, 1G networks. These networks were designed in the mid 80s to deliver voice services to

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

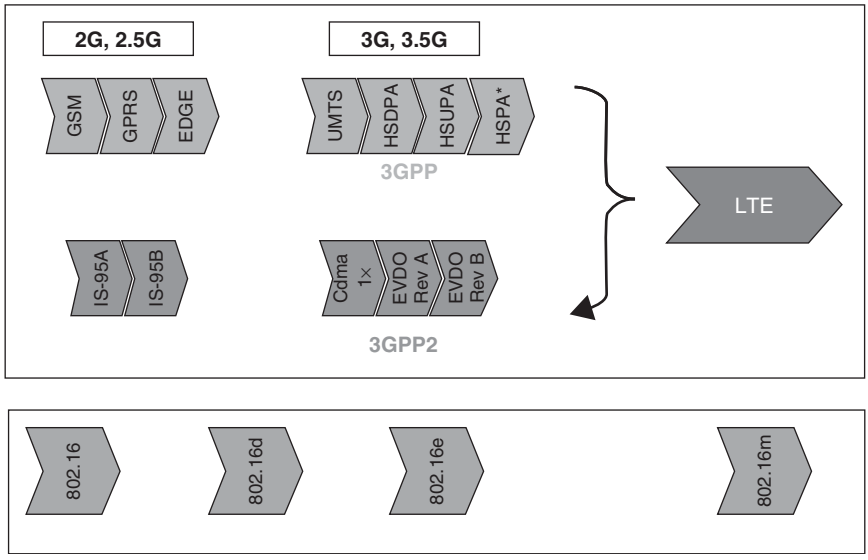


Figure 15.1 LTE and WiMAX Evolution.

mobile users. Shortly afterwards, the 2G circuit-switched networks had evolved to improve the quality of these services. This has been achieved through using digital communications techniques, including modulation and coding. In this generation, two technologies prevailed, the TDMA based GSM and the CDMA based IS-95. The development of these two technologies was led by two standardization bodies, 3GPP leading GSM and 3GPP2 leading IS-95. GSM supported voice services neatly; however, it fell behind in data services. This motivated the development of GPRS technology as a 2.5G network. This technology could successfully handle data services over the GSM circuit switched network. Since then, 3GPP continued its work in improving the performance of the data transmission capabilities by increasing the supported data rates over the GSM networks. This led to the introduction of EDGE in 2003, a technology that is capable of providing a theoretical data rate of 1 Mbps – three times the data rate of GSM networks.

The next step in this development line was the introduction of HSPA, a 3G technology capable of delivering theoretical data rates up to 14 Mbps. The data rate was improved even further with the introduction of HSPA+ that increased the data rates over the two link directions; the uplink and the downlink. This technology is considered the big step towards LTE. It can be considered as a revolutionary development along the line, since it departed from the split circuit-packet switch network by introducing, for the first time, an all-IP architecture as an option for the voice and data services. In addition, HSPA+ integrated the MIMO technology as a major part of its PHY layer, paving the way for its integration in LTE.

A similar line of evolution was followed by the 3GPP2. It has progressively evolved IS-95 form a mere voice services network into the Evolution Data Optimized (EVDO) Rev B, a network that efficiently supports both data and voice services and at various mobility levels.

Meanwhile, the standard for WiMAX networks was produced by the IEEE in an attempt to extend the WiFi-like services into metropolitan and wide area networks, but at much higher data rates. Earlier work on the IEEE 802.16 standard was largely based on the Data Over Cable Service Interface Specification (DOCSIS), modifying the MAC layer to be more suitable to the wireless interface. In 2004, IEEE modified this version by introducing OFDM. This was the first version of the IEEE 802.16d standard. Hence, IEEE 802.16 originated from computer networks that are inherently efficient for data services. In fact, this is one of the differences between LTE and IEEE 802.16x. While LTE is a revolution of the voice services network, IEEE 802.16 is a revolution of the data services network.

The main motive for the development of the two standards was hence the need to offer higher and more reliable data rates to accommodates the increasing demand for mobile data traffic. Additionally, customers are expecting the service provider networks to support several types of applications including the bandwidth-hungry applications such as video streaming, video conferencing and gamming. Coping with the increased number of users and providing the continuous support for current and new developed bandwidth-hungry applications leave the broadband wireless communication designers restless finding different means to provide higher data rates. Figure 15.2 shows the gap between the available capacity of current technologies and the demand for data rate.

ITU produced a new set of requirements for future wireless networks. These requirements mainly focus on providing high data rates. In particular, it specifies

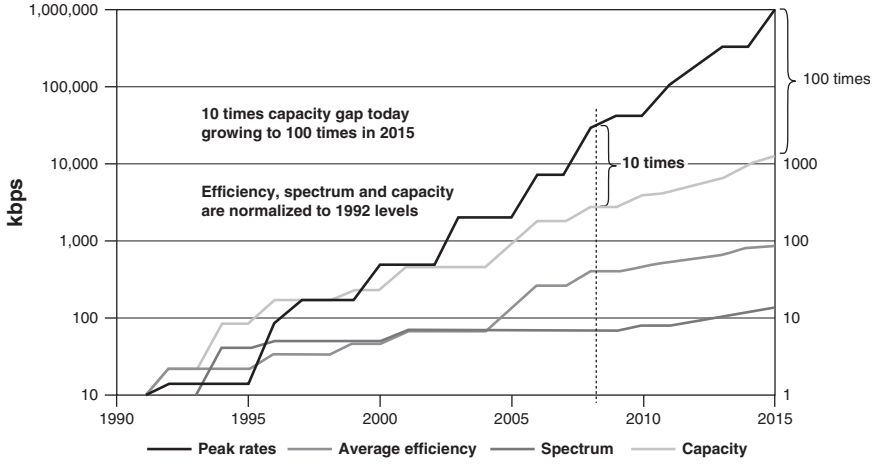


Figure 15.2 The gap between the capacity and demand in bps.

a 1Gbps for fixed and slow mobility users and a 100Mbps for high mobility users. Neither IEEE 802.16-2009 nor LTE Rev.9 is capable of providing such rates. Hence, the two standardization groups, the 3GPP and the IEE needed to come with its own IMT-Advanced proposal. While the former proposed LTE-Advanced, the latter proposed IEEE 802.16m. As the names suggest, these two proposals are in fact enhancements of their ancestors, the LTE-Rev.9 and the IEEE 802.16-2009, respectively.

15.2 Comparing Spectral Efficiency

A thorough inspection of the spectral efficiencies of the two technologies on the UL and the Dl unveils a close similarity. These efficiencies are summarized in Table 15.1.

In order to achieve these high spectral efficiencies, the two proposals tend to utilize the latest technology advances at the PHY layer. These include multicarrier communication (OFDM, SC-FDMA, and OFDMA), adaptive MIMO with up to four layers, flexible spectrum and fractional frequency reuse, relaying, multi-cell MIMO, etc.

15.2.1 OFDMA Implementation

OFDMA, as a multi access technology, offers a significant improvement in the spectral efficiency for more than one reason. For instance, its inherent multi-path interference handling capability facilitates delivering high data rates while experiencing marginal ISI. Also, the possibility of allocating different sizes of bandwidth chunks for users helps providing different data rates and accommodating more users. Moreover, OFDMA integrates smoothly with MIMO technology, which is another rate-boosting technology.

In WiMAX, OFDMA is adopted for the DL as well as the UL. However, LTE resorted to SC-FDMA on the UL to enhance the power efficiency of the MS. This choice, as the 3GPP argues, reduces the PAPR on the UL by 1–2 dB, hence prolonging the lifetime of the battery. In fact, WiMAX adopted an alternative approach to achieve a similar reduction in the PAPR. It depends on designing

Table 15.1 A comparison between the spectral efficiency performance of LTE and WiMax

	LTE		WiMAX	
DL Spectral Efficiency	1.57 bps/Hz/Sector (2 × 2) MIMO	1.59 bps/Hz/Sector (2 × 2) MIMO	30 bps/Hz	>2.6 bps/Hz (4 × 2)
UL Spectral Efficiency	0.64 bps/Hz/Sector (1 × 2) SIMO	0.99 bps/Hz/Sector (1 × 2) SIMO	15 bps/Hz	>1.8 bps/Hz (2 × 4)

efficient resource allocation schemes. However, such schemes render the UE design process more complicated.

Even though the two networks utilize OFDMA, they differ in its implementation, especially in implementing the frame structure. LTE and LTE-Advanced use a fixed frame size of 10 ms with a subframe size of 1ms. Meanwhile, IEEE 802.16-2009 defines a frame size of variable duration, 2 to 10 ms). The IMT-Advanced IEEE 802.16m uses a hierarchical frame size with unit frames of duration 5 ms. The driver for that latter choice was to facilitate backward compatibility in IEEE 802.16-2009. Generally, however, the shorter frame duration in the IEEE 802.16m amendment was need to meet the ITU-R delay requirements.

One of the main advantages of OFDMA is the flexible resource allocation. In LTE, resources are allocated every subframe, with control messages broadcasted in the first three DL OFDM symbols. On the other hand, WiMAX allocates the resources every frame. However, since WiMAX adopts variable frame sizes, the duration of the scheduling cycle depends on the frame size. At least two DL OFDM symbols are allocated for control messages.

Figure 15.3 shows the difference in implementing OFDMA between LTE and WiMAX. In LTE, 12 contiguous subcarriers are grouped to form a RB in frequency domain and either six or seven ODFM symbols, depending on whether the normal or extended cyclic prefix is employed, in the time domain. The OFDMA subcarriers are spaced by fixed frequency of 15 KHz in the frequency spectrum. The number of RBs in a channel varies between six and 100 depending on its bandwidth. WiMAX allows contiguous subcarrier grouping called Band Adaptive Modulation and Coding (BAMC) to form a time slot, or a distributed subcarrier grouping known as PUSC (see Chapter 4). In PUSC, 24 subcarriers are grouped in frequency domain over two consecutive OFDM symbols producing one time slot of 48 subcarriers, while BAMC groups 16 subcarriers over three consecutive OFDM symbols resulting in 48 subcarriers. Irrespective of the sub-channelization method used, a WiMAX slot is always formed from 48 subcarriers. In both grouping methods, the number of subcarriers per slot is 48. Unlike LTE, WiMAX does not specify a fixed subcarrier spacing. The frequency guard band depends on the channel bandwidth; in case of a 20 MHz bandwidth the frequency guard band employed by WiMAX is 10.94 MHz.

15.2.2 MIMO Implementation

MIMO is another example of rate-boosting PHY techniques adopted in by the two technologies. Unlike OFDMA, MIMO enhances the data rate without any increase in the channel bandwidth. However, similar to OFDMA, LTE and IEEE 802.16-2009 have different ways in implementing MIMO.

LTE adopted several MIMO techniques including SU-MIMO, MU-MIMO, open-loop and closed-loop spatial multiplexing, and dedicated beamforming. SU-MIMO is supported in the DL with up to four layers while MU-MIMO

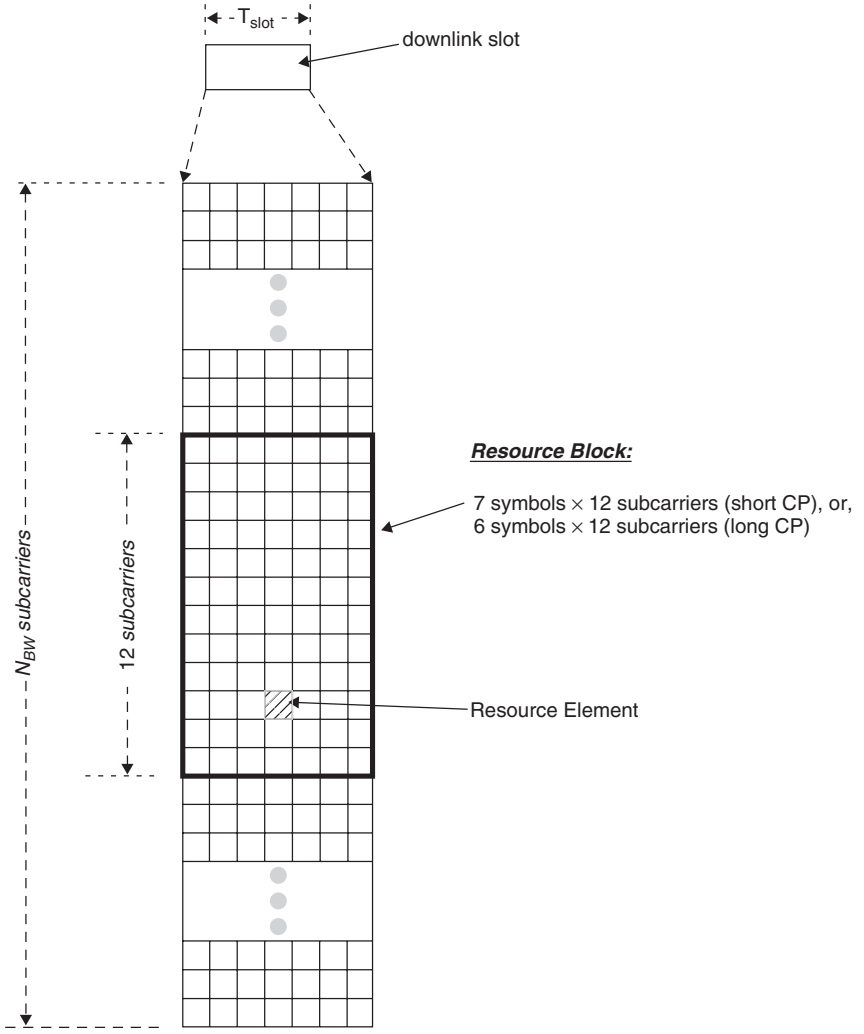


Figure 15.3 The OFDMA Frame Structure.

is supported in both, the UL and the DL, with up to four layers in the DL and two layers in the UL.

IEEE 802.16-2009 adopts two open loop MIMO techniques, namely MIMO Matrix-A (Space Time Block Coding) and MIMO Matrix-B (Spatial Multiplexing). The IEEE 802.16e standard also includes two and four antenna MIMO systems; however, its application focuses on 2×2 antennas. On the other hand, IEEE 802.16-2009 enhanced the data rate by increasing the number of antennas at the terminal. Additionally, the standard introduced closed-loop codebook-based precoding only for the TDD mode of operation.

Compared to LTE and IEEE 802.16-2009, LTE-Advanced and IEEE 802.16m underwent MIMO technology enhancements. Both networks introduced the MU-MIMO, where more than one mobile user can be assigned to one resource block at the same time, and increase the number of DL and UL transmissions to eight and four, respectively. This enhances the spectral efficiency and increases the data rate as per IMT-Advanced requirements. Multi-cell or network MIMO is another MIMO advancement utilized in LTE-Advanced and IEEE 802.16m. In it, multiple BSs collaborate to serve multiple MSs at the cell edge. Multi-cell MIMO eliminates inter-cell interference and provides diversity gains, which improve the data rates of cell-edge users, consequently, increase the whole cell average throughput.

CoMP is considered for LTE-Advanced as a tool to improve the cell-edge throughput. A CoMP is a system of several BSs distributed over a certain geographical area connected to each other over a dedicated link. The BSs coordinate transmission/reception of a particular user to enhance the communication reliability. CoMP can be of two types: dynamic scheduling between the multiple cells in the geographical area and joint transmission/ reception from the multiple cells. Dynamic scheduling is not a novel concept for LTE-Advanced. It can be viewed as an extension to inter-cell interference management supported in LTE. In inter-cell interference management, scheduling is coordinated among multiple neighboring BSs in order to achieve adaptive inter-cell interference coordination. Joint transmission is achieved by having multiple BSs transmitting to a single user. Through this, interference is reduced and the received power is increased. Both CoMP and Multi-cell MIMO in LTE-Advanced and IEEE 802.16m respectively improve throughput and converge the user's mobility experience.

15.2.3 Spectrum Flexibility

IEEE 802.16 and LTE are distinguished from 2G and 3G networks by their scalable spectrum allocations. 2G and 3G networks are defined over a fixed width spectrum. Meanwhile, IEEE 802.16-e, which is limited to the TDD duplexing mode, can operate over the 2.3, 2.5 and 3.5 GHz licensed bands and the 5.3 unlicensed band. IEEE 802.16-2009 added two more spectrum bands, 1.7 and 2.1 GHz, mostly to accommodate FDD. To comply with the IMT-Advanced frequency bands, IEEE 802.16m is defined in a new set of spectrum bands as shown in Table 15.2. The frequency bands for LTE are summarized in Tables 15.3 and 15.4.

LTE and IEEE 802.16-2009 are both expected to support different types of applications with diverse QoS requirements. Data rates achieved by LTE and IEEE 802.16-2009 depend on channel bandwidth, number of MIMO layers used and modulation type. IEEE 802.16-2009 defines different channel bandwidths from 5 MHz–28 MHz. In addition to these, IEEE 802.16m defines an optional channel bandwidth of 40 MHz without carrier aggregation, with which, channel bandwidth can be increased up to 100 MHz. LTE defines 1.4, 3, 5, 10, 15, 20 MHz channel bandwidths, which can be increased up to 100 MHz using spectrum

Table 15.2 IEEE 802.16m Frequency Band

Bond Class	UL AMS Transmit Frequency (MHz)	DL AMS Receive Frequency (MHz)	Duplex Mode
1	2300–2400	2300–2400	TDD
2	2305–2320, 2345–2360	2305–2320, 2345–2360	TDD
	2345–2360	2305–2320	FDD
3	2496–2690	2496–2690	TDD
	2496–2572	2614–2690	FDD
4	3300–3400	3300–3400	TDD
5L	3400–3600	3400–3600	TDD
	3400–3500	3500–3600	FDD
5H	3600–3800	3600–3800	TDD
6	1710–1770	2110–2170	FDD
	1920–1980	2110–2170	FDD
	1710–1755	2110–2155	FDD
	1710–1785	1805–1880	FDD
	1850–1910	1930–1990	FDD
	1710–1785, 1920–1980	1805–1880, 2110–2170	FDD
	1850–1910, 1710–1770	1930–1990, 2110–2170	FDD
7	698–862	698–862	TDD
	776–787	746–757	FDD
	788–793, 793–798	758–763, 763–768	FDD
	788–798	758–768	FDD
	698–862	698–892	TDD/FDD
	824–849	869–894	FDD
	880–915	925–960	FDD
	698–716, 776–793	728–746, 746–763	FDD
8	1785–1805, 1880–1920, 1910–193, 2010–2025, 1900–1920	1785–1805, 1880–1920, 1910–193, 2010–2025, 1900–1920	TDD
9	450–470	450–470	TDD
	450.0–457.5	462.5–470.0	FDD

aggregation. Table 15.5 shows the data rates of LTE and IEEE 802.16-2009 at a channel bandwidth of 20 MHz and different MIMO layers. Table 15.6 shows the data rates of LTE-Advanced and IEEE 802.16m.

As mentioned earlier, increasing the transmission bandwidth is one of the many ways through which data rates can be enhanced. In fact, supporting wideband transmissions of up to 40 MHz (more is encouraged but not required) is one of

Table 15.3 LTE–FDD Frequency Band Allocations

Band Number	Band Description/ Name	Uplink (MHz)	Downlink (MHz)
1	IMT core	1920–1980	2110–2170
2	PCS 1900	1850–1910	1930–1990
3	GSM 1800	1710–1785	1805–1880
4	AWS (US)	1710–1755	2110–2155
5	850 (US)	824–849	869–894
6	850 (Japan)	830–840	875–885
7	IMT Extension	2500–2570	2620–2690
8	GSM 900	880–915	925–960
9	1700 (Japan)	1749.9–1784.9	1844.9–1879.9
10	3G Americas	1710–1770	2110–2170

Table 15.4 LTE-TDD Frequency Band Allocations

Band Designation	Band Name	Allocation (MHz)
a	TDD 1900	1900–1920
b	TDD 2.0	2010–2025
c	PCS centre gap	1910–1930
d	IMT extension centre gap	2570–2620

Table 15.5 Data rates of LTE and IEEE 802.16-2009 at 20 MHz channel bandwidth and different MIMO settings

Parameter	Reported LTE Results				WiMAX Rel 1.5	
	Motorola		T-Mobile	Qualcomm		
BS Antenna	2 × 2	4 × 4	2 × 4	4 × 2	2 × 2	4 × 4
Channel BW	2 × 20 MHz				2 × 20 MHz	
DL Peak User Rate	117 Mbps	226 Mbps	144 Mbps	277 Mbps	144.6 Mbps	289 Mbps
MS Antenna			1 × 2	1 × 2	1 × 2	
UL Peak User Rate	N/A	N/A	50.4 Mbps	75 Mbps	69.1 Mbps	

Table 15.6 Data rates of LTE-Advanced and IEEE 802.16m

Parameter	LTE-Advanced	IEEE 802.16m
DL Peak User Rate	1 Gbps	DL >350 Mbps (4 × 4) @ 20 MHz FDD
UL Peak User Rate	300 Mbps	UL >200 Mbps (2 × 4) @ 20 MHz FDD

the IMT-Advanced requirements. However, such an increase is not easily achievable especially with the contemporary spectrum scarcity and with the backward compatibility requirement (i.e., with LTE and IEEE 802.16-release 2). Nevertheless, wider transmission bandwidths can still be achieved by either spectrum aggregation or dynamic spectrum allocation.

Spectrum aggregation is supported in IEEE 802.16m and in LTE-Advanced. In LTE-Advanced, two or more RF carriers, each with a bandwidth up to 20 MHz can be aggregated, reaching a maximum bandwidth of 100 MHz. In IEEE 802.16m, on the other hand, the bandwidth can be from 5–40 MHz. While spectrum aggregation enables access to large allocations (up to 100 MHz), it requires contiguous free allocations to be made. This availability may not always be possible. In this case, the aggregation of noncontiguous spectrum chunks becomes desirable.

The second mechanism that facilitates achieving wide transmission bandwidths is dynamic spectrum allocation. This is particularly the case in IMT-Advanced networks that will be able to benefit from spectrum not previously assigned to IMT systems. Such allocations also allow for flexible spectrum usage among several operators, facilitating better radio resource management and allow offering services with higher data rate. Even though, both candidate technologies support carrier aggregation and dynamic spectrum allocation, IEEE 802.16m compared to LTE-Advanced faces the lack of high quality spectrum. The available spectrum for IEEE 802.16m is still limited to 3.5 GHz or 5 GHz, which is higher than that of LTE-Advanced.

15.3 Comparing Relay Adoption

Wireless multihop relaying entails delivering an MT's connection to the BS through dedicated RSs. Both candidate technologies show interest in introducing and enhancing relayed transmission. LTE-Advanced defines two types of relay: Type-I and Type-II, while IEEE 802.16j defines transparent and non-transparent relaying. The main objective of relaying in IMT-Advanced systems is to extend the cell coverage through the RS and enhance the overall cell throughput. Type I and non-transparent RSs extend the BS's coverage to include MTs that cannot connect directly to the BS. Such RSs are required to broadcast control information to the MTs as the MTs cannot receive the BS's own control transmission.

Meanwhile, Type II RSs in LTE-Advanced and transparent RSs in IEEE 802.16m are not used for extending a BS's coverage. Rather, they are used to enhance the service quality and the link capacity within the cell coverage area. These RSs do not need dedicated control messages as the MTs can identify the BS's own control messages.

Relaying aids in meeting the user requirements from three perspectives: increased coverage, higher throughput and improved reliability. Through relaying, a user can easily roam over considerably longer distances with the support of the same network technology. In areas with strong fading, a RS enhances network connectivity and reliability, and extends its coverage. However, Type II and transparent relaying are used to realize higher throughput and data rates, supporting multimedia applications and providing for their QoS requirements.

Little difference can be noted between the Type II RSs and the transparent RSs. There are, however, apparent differences between the Type I RSs and the non-transparent RSs. Type I RSs are limited to two hops, while non-transparent RSs are unlimited in the number of hops. This limitation, reflected in the relevant designs of frame structures and signaling, gives an advantage to IEEE 802.16m (which is based on IEEE 802.16j for multihop communication) over LTE-Advanced. IEEE 802.16m can widen the coverage area in a cost-efficient manner. However, it should be remarked that multiple hops results in substantial increase in the delay.

15.4 Comparing Network Architectures

Conceptually, LTE and WiMAX have similar network architectures. Both have an all-IP flat architecture. Their network architectures can be divided into three logical parts: Mobile Station (MS)/User Equipment (UE), Access Service Network (ASN)/Core Network (CN) and Connectivity Service Network (CSN)/Protocol Data Network (PDN). Figure 15.4 and Figure 15.5 show the architectures of WiMAX and LTE respectively. The two networks differ in the functionalities performed by the first part, but not its architectural aspect. Hence, in the sequel, we shall highlight the differences between the two technologies in the last two parts.

15.4.1 ASN/AN (E-UTRAN) and the MME and the S-GW

The ASN in WiMAX consists of an ASN Gateway (ASN-GW) and a BS. The AN (EUTRAN) consists of a network of eNBs connected to each other. The BS is functionally similar to an eNB. The main task of the two is handling traffic to and from the MS. This involves packet transmission, HARQ, link adaptation, and QoS enforcement at the user plane. At the control plane, it involves radio resource management, connection management, handover triggering and DHCP proxy at the control plane.

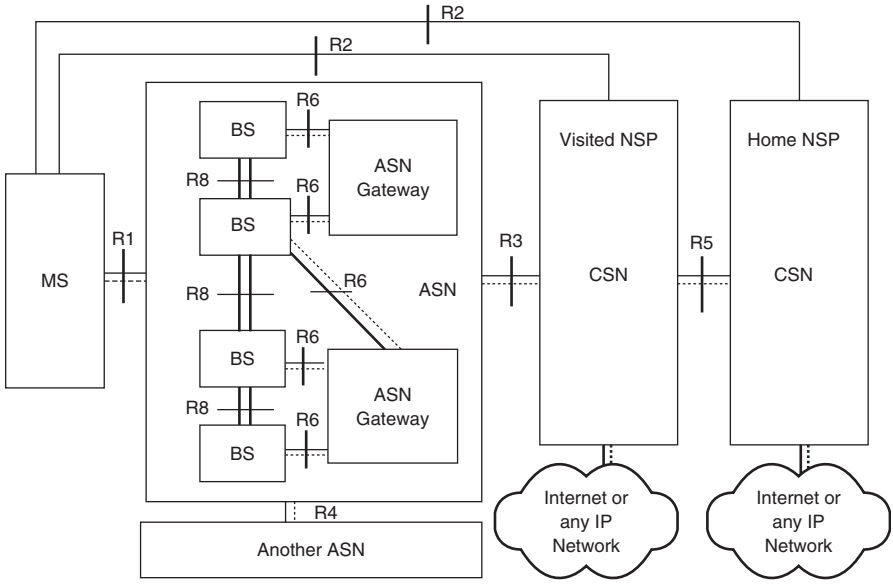


Figure 15.4 WiMAX Network Architecture.

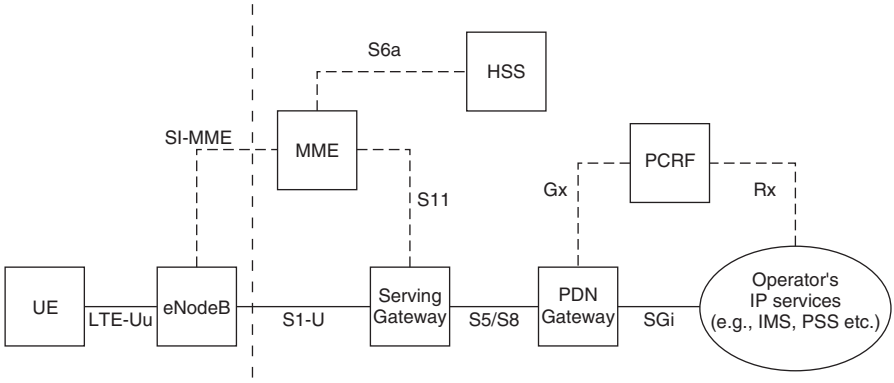


Figure 15.5 LTE Network Architecture.

The gateway function played by the ASN-GW in the WiMAX is provided by two entities in LTE which are part of the EPC, namely the MME and the S-GW. These two are functionally similar to the SGSN in UMTS or the PDSN in EVDO.

The ASN-GW in WiMAX as well as the MME and the S-GW in LTE provide the air-interface to the core network and are considered the aggregation point of the users' traffic. Among the tasks of these entities are the AAA client

procedures, mobility management by establishing, maintaining and terminating mobility tunnels with the eNB/BS.

Irrespective of the similarities in functionality between WiMAX and LTE, significant differences exist. For example LTE separates the traffic control plane handling from the traffic user plane handling. MME is defined to handle the control plane traffic while the S-GW is defined to handle the user plane traffic. WiMAX does not separate the two planes at least in definition, where the control and the user planes' traffic are handled by the ASN-GW. In addition, LTE defined entities (part of MME and S-GW) to provide interface between LTE and legacy 3G networks such as WCDMA and EVDO. However, WiMAX is still working on this unfinished functionality.

15.4.2 CSN/PDN-GW

CSN provides connectivity to other networks such as Internet and the PSTN (Public Switched Telephone Network). The main task of CSN is to provide core IP functionality and an anchor for mobility (Mobile IP Home Agent MIP-HA) from WiMAX to/from other network technologies. The PDN-GW provides similar mobility functionality to the HA of the CSN. Concerning mobility functionality, LTE provides proxy MIPv6 while IEEE-IEEE 802.16-2009 left this functionality as an option.

Both IEEE 802.16m BS and LTE-Advanced eNB are standardized to support interoperability with IEEE 802.16-e and IEEE 802.16-2009 and LTE respectively. Additionally, LTE-Advanced eNB and IEEE 802.16m BS will be capable to serve legacy UE and MS.

16

Coexistence and Inter-Technology Handovers

Coexistence is defined as "The ability of one system to perform a task in a given shared environment where other systems have an ability to perform their tasks and may or may not be using the same set of rules" [1]. In the context of wireless communications, the IEEE 802.19 Wireless Coexistence Working Group (WG) defines it as the simultaneous use of the same spectrum in the same locale by two or more wireless devices or the act of two or more networks/devices sharing resources without causing destructive interference to one another [2]. The IEEE 802.19 WG develops standards for coexistence between wireless standards operating in unlicensed bands. As multiple wireless network technologies either already exist or are currently in development for near future deployment, seamless coexistence is becoming a more challenging objective to achieve. Meanwhile, the IEEE 802.21 WG oversees the Media Independent Handovers (MIH), which dictates the procedures for IEEE technologies when managing inter-technology handovers.

This chapter discusses the coexistence from the point view of LTE as an example of coexistence between LTE or WiMAX with other wireless access systems and fixed (such as satellite services) and mobile (such as IEEE 802.11 networks). It is organized into two sections. The first, Section 16.1, discusses intersystem interference, and shows an example of how it can be managed, while Section 16.2 discusses inter-technology handover.

16.1 Intersystem Interference

Intersystem interference is the induced unwanted power in a communication system made by other communication systems using the same frequency band(s). Nowadays, advanced communication devices can be equipped with more than one

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

radio access interface to allow it to connect to a number of access technologies. Examples of these access technologies are GSM, (E)GPRS, enhanced EDGE, UMTS, HSPA, evolved HSPA, LTE, WiFI, and WiMAX. While multi-technology handsets enable an opportunistic, user-centric service access, this enabled feature comes at the price of frequency interference that mainly occur at the mobile handset. In a multi-access system, the user device may undergo or cause substantial interference between radio transceivers of different access technologies. Intersystem interference can be alleviated, even mitigated, if a priori knowledge is available about the specific access technologies operating in a shared area, in addition to the bands that they utilize. A difficulty emerges, however, in that such a priori knowledge implies fixed frequency allocation, as is common with 2G/3G technologies. LTE and LTE-Advanced, however, have been empowered with greater flexibility in having a large number of radio spectrum allocations and for both duplexing types (FDD and TDD). This flexibility renders achieving a prior knowledge about the exact LTE spectrum bands too difficult. The emerging usage of cognitive radios introduces another source of uncertainty as to the sources of intersystem interference, wherein flexible spectrum allocation of any frequency band is feasible.

To mitigate this interference, efficient access mechanisms are required to identify the network to which a user should be connected in a multi-access system. The priority to connect to the user's preferred network must be identified, the level of interference with other technologies in the vicinity using the same frequency band need to be recognized and the transmission and reception requirements must be known.

16.1.1 Types of Intersystem Interference

Intersystem interference can be due to either in-band or out-of-band emissions. In-band emissions result in causing interference to the victim system. Out-of-band emissions are frequency components that are outside the operational bandwidth but that interfere due to the utilized modulation process and unintentionally amplified harmonics. They result from hardware limitations and non-linearity of the radio frequency transceivers, both of which cause unintended RF emissions beyond the designated bandwidth. Although the impact of such emissions can be reduced by applying more stringent requirements for filtering broadband wireless signals, these requires would complicate, and hence increase the production cost, of broadband wireless equipment. A third cause of intersystem interference is in signals from nearby secondary broadband wireless transmitters causing receiver saturation of the victim systems.

The protection of the victim system from interference can be achieved by means of insuring minimum separating distances between LTE and other wireless systems. This solution is mainly applicable for fixed satellite services (FSS). Recommended distance is calculated based on different studies carried out to evaluate the impact of IMT-Advanced technologies intersystem interference on the normal operation of FSS systems. To mitigate in-band co-frequency

emissions, the report in [3] recommends a distance separation in tens of kilometers. If such separation is not feasible, site shielding of the broadband wireless system (for fixed broadband wireless stations) must be applied. For protection against out-of-band emissions, a distance of 2 km is required. An alternative to this separation would be applying additional filtering at the broadband wireless stations to reduce out-of-band emissions to a level not exceeding -89dBw/MHz. In case of Receiver Saturation problem, the separation distance can be 0.5 – 0.6 Km without filtering and 2 Km with filtering.

A solution for coexistence of LTE-TDD with IEEE 802.16m-TDD is proposed in [4] by having the LTE TDD portion and the IEEE 802.16m portion sharing the IEEE 802.16m air link in a time-division manner. In [4], the frame structure of IEEE 802.16m is proposed to be changed as shown in Figure 16.1 in order to facilitate the in-band coexistence of LTE-TDD and as a solution for the co-frequency emission problem.

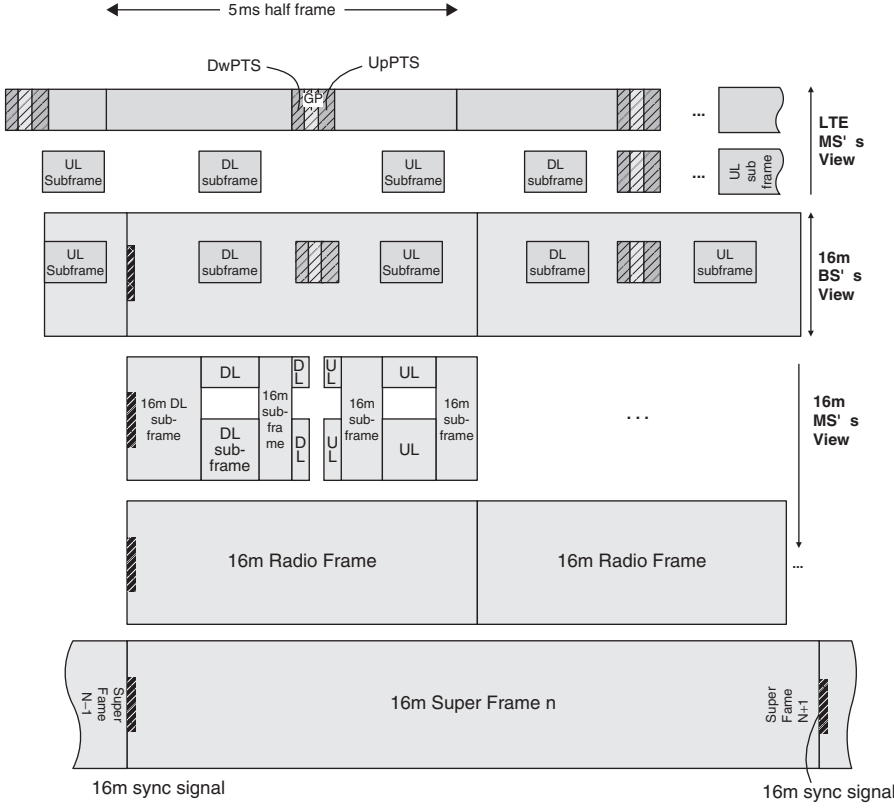


Figure 16.1 Frame Structure Supporting LTE-TDD with IEEE 802.16m.

16.2 Inter-Technology Access

Inter-radio access technology is the ability to support the mobility of a user device between differing radio access network types, also known as vertical handovers. Inter-radio access between heterogeneous networks is important to be executed properly for seamless communication with minimum delay and packet loss to disassociate from the current serving BS to associate to the new BS. LTE defines two types of inter-radio technologies: inter- technology access between LTE to legacy 3GPP systems, also known as inter-radio access technology (inter-RAT) and inter- technology access between LTE and other non-3GPP systems.

Inter-technology access can be supported using different techniques. The most primitive one is the mobility from one technology to another without the intervention of the network. In this case the device is equipped with different technology interfaces. The user or the device selects which technology to access and associate to the corresponding network. Once this network becomes unavailable, the user or the device selects another technology and associate with. This type of inter-technology handover is acceptable for delay-tolerant and low QoS requirement applications such as http and e-mail. However, delay-intolerant or session based applications cannot tolerate the service interruption and may require re-initiation of the session including the re-authentication process. A more efficient inter-technology mobility for session based applications is one that supports the data session continuity across multiple technologies. In this type of mobility, session continuity is preserved while the user moves across different technologies. Association to a new technology and disassociation from a serving one is accomplished with no user actions and it is transparent to the applications. Hence, re-authentication by users or data interruption has no impact in this type of mobility.

16.2.1 Approaches to Inter-Technology Mobility

Different approaches are used to provide inter-technology mobility with session continuity [5]. We discuss three general approaches in this section:

Single Transmit Device: Mobile IP (MIP) is standardized by the IETF to support for the session continuity at the IP layer. Hence, cannot support user authentication and login while moving across the different technologies. Single transmit device- MIP based approach makes use of the MIP service and hence it is a single transmit device, the device is only capable to associate with one technology at a time. In other words, it needs to disassociate from the serving technology before associating to the target technology. Despite the simplicity of this type, it suffers from a large delay associated with the signaling needed to associate and authenticate with the target technology. The type of inter-technology handover is known as non-optimized inter-technology handover. 3GPP standard use this approach to support inter-technology mobility between WiMAX and LTE and between

EVDO and LTE. Optimized inter-technology handover is defined in LTE as the inter-technology handover which allows/requires the serving technology exchanging control data and signaling messages with the target technology as described in the following approach. The optimized handover is expected to support delay-stringent applications such as VoIP.

Access Network Interconnect: This approach is used mainly in networks managed by single operator and employ technology with different generations but similar origin (newer technology that supports backward compatibility with older generations) such as CDMA2000 and EVDO or UMTS and GSM. The access network interconnect requires the serving and the target networks to be physically connected to facilitate the handover process and exchange the signaling messages (optimized handover). Access network interconnect is limited to the technologies produced by the same standardization body, however, 3GPP with collaboration with 3GPP2 defined procedures to extend this approach to EVDO. 3GPP and IEEE 802.16 working group are in the process in investigating optimized handover between LTE and WiMAX.

Dual-Transmit Devices (DTD) based inter-technology handover: This type of mobility does not require the serving technology to be connected to the target technology, since the user device is involved in the initiation and termination of the connection to the technology. It is realized by two types of services, Mobile IP (MIP) and Session Initiation Protocol (SIP).

MIP: In this type, since the device is equipped with dual transmitters, the device during handover employs make-before-brake handover. The device while it is connected to its serving technology uses its second transmitter to connect to the target technology, hence maintaining its data session and preventing data loss. Once the association process is completed with the target technology, the device uses the MIP service to move the data session to the target network. Examples are the inter-technology handover between LTE and WiFi and EVDO and WiMAX.

SIP: This solution is suitable for inter-device inter-technology mobility where a data session is required to be moved not only between technologies but also between devices. This is the only solution to support inter-device mobility. However, it is only applicable for SIP based applications. An example of this type of mobility is the standardization of LTE-Advanced which is expected to support inter-device and inter-technology mobility based on SIP and IMS.

16.2.2 Examples of Inter-Technology Access

16.2.2.1 Inter-RAN Between LTE and CDMA2000

We present in this subsection, the support of the mobility between LTE and CDMA2000 as an example of the Inter-technology access between 3GPP2 and LTE system. Figure 16.2 shows the network architecture support for the

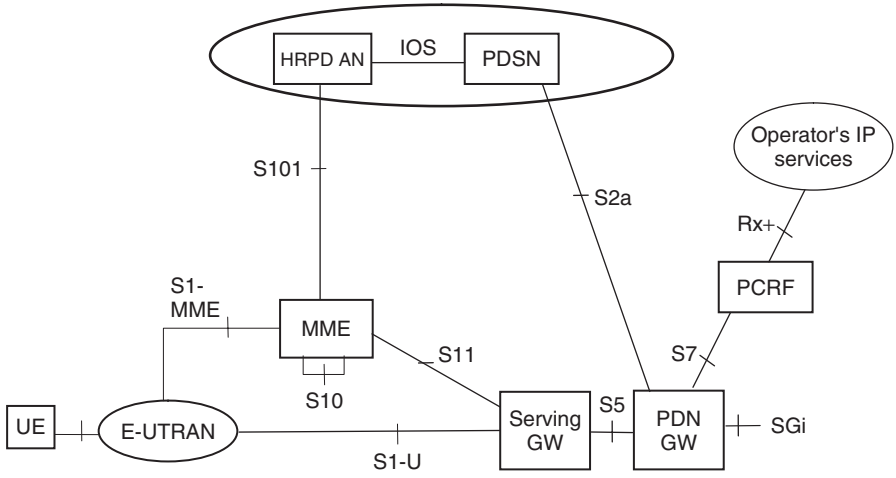


Figure 16.2 Architecture for optimized handover between mobile WiMAX and 3GPP2.

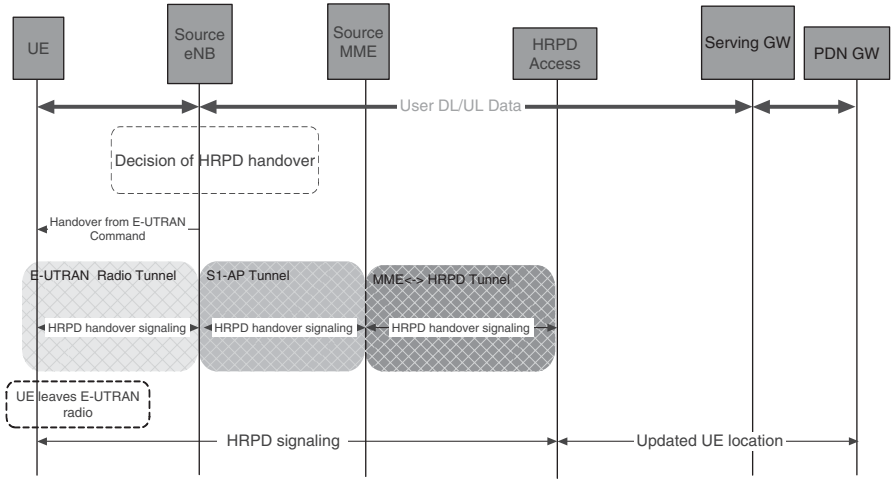


Figure 16.3 Handover procedure from LTE to CDMA2000.

mobility between CDMA2000 and LTE, and Figure 16.3 shows the procedure for the handover between LTE and CDMA2000.

A UE is attached to the EUTRAN network. Based on measurement reports received from the UE, the eNB initiates a handover by sending a "Handover from EUTRAN Command" message to the UE to indicate that the UE should begin the handover procedure. The message includes the specified target type and any

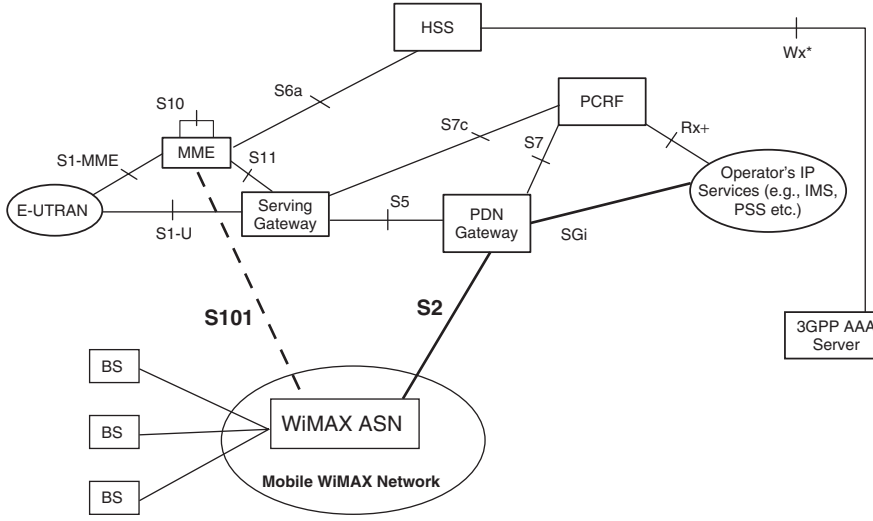


Figure 16.4 Architecture for optimized handover between mobile WiMAX and 3GPP using L2 Tunneling.

specified parameters needed by the UE to create the appropriate messages needed to request a connection from the CDMA2000 network. The UE continues to send and receive data on the EUTRAN radio until it receives the "handover command" ordering it to switch to the target CDMA2000 cell. After the "handover command" is received by the UE, it leaves the EUTRAN radio and start acquiring the CDMA2000 traffic channel. When the UE receives the CDMA2000-HRPD Traffic Channel Assignment Message (tunneled over the EUTRAN), it leaves the EUTRAN radio and perform its access over the CDMA2000-HRPD radio.

Figure 16.4 shows the reference architecture for optimized handovers between mobile WiMAX and 3GPP access using L2 tunneling between MME and WiMAX ASN. This architecture uses the EPC network elements and reference points which are already specified. It does not require any changes on these network elements and reference points. All the interfaces and network entities that separate SAE/LTE from WiMAX ASN are defined in [4]. Interface S101 enables interaction between EPS and WiMAX ASN access to allow for handover signaling.

16.2.2.2 Inter-RAN Between LTE and WiMAX

LTE release 9 defines the procedure for optimized LTE to WiMAX handover. The following steps show the procedure for a UE to move from an LTE network to a WiMAX as presented in [6]. These steps are shown in Figure 16.5.

Based on the Measurement Report received from the UE, EUTRAN may trigger the UE to perform WiMAX measurements. Configurations for WiMAX

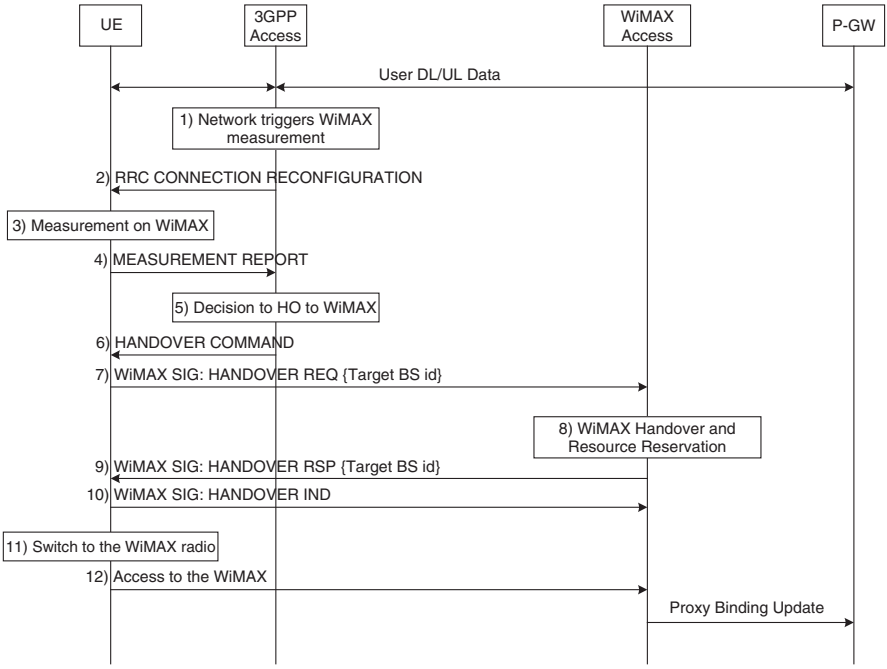


Figure 16.5 Optimized EUTRAN to WiMAX Handover.

measurements and measurement reports are sent to the UE. This latter performs measurement on the WiMAX based on the received WiMAX measurement configuration. Next, it sends Measurement Report based on the received WiMAX measurement reporting configuration. Based on the Measurement Report received from the UE, EUTRAN may decide handover to the WiMAX for the UE. EUTRAN may also decide handover based on RRM information.

EUTRAN instructs the UE to initiate handover to the WiMAX by Handover Command. EUTRAN can inform whether optimized handover is supported or not. If optimized handover is supported, steps 7–12 will be followed. If optimized handover is not supported, after the reception of the Handover Command, the UE will leave the 3GPP radio access, switch to the WiMAX radio access, and perform WiMAX specific handover procedure. The UE initiates the handover to the WiMAX by tunneling a WiMAX Handover Req. message including the target WiMAX BS ID. Resources are reserved in the target WiMAX. The WiMAX sends a WiMAX Handover Rsp. message including the target WiMAX BS ID. The UE notifies the WiMAX that it starts handover to the indicated WiMAX BS by tunneling a WiMAX Handover Ind. message.

The UE leaves the 3GPP radio access and switches to the WiMAX radio access. The UE performs the WiMAX specific access procedure.

References

[1] 16.2-2003, Part 16.2: Coexistence of Wireless Personal Area Networks with Other Wireless Devices Operating in Unlicensed Frequency Bands, 28 August 2003.
[2] IEEE 802.19, SDD-Terminology-Strawpoll-Results, 16, March 2010.
[3] Asia-Pacific Telecommunity, "Report on co-existence of broadband wireless access networks in the 3400–3800 mhz band and fixed satellite service networks in the 3400-4200 MHz band", The 3rd Interim Meeting of the APT Wireless Forum Document AWF-IM3/10 (Rev.1), January 2007, Bangkok, Thailand.
[4] IEEE 802.16 Broadband Wireless Access Working Group, "Proposal for IEEE 802.16m TDD Coexistence with LTE TDD on a Co-channel Basis", 2008.
[5] Motorola, "LTE Inter-technology Mobility Enabling Mobility Between LTE and Other Access Technologies".
[6] 3GPP TR 36.938-900: "Improved Network Controlled Mobility between EUTRAN and 3GPP2/Mobile WiMAX Radio Technologies (Release 9)".

17

Supporting Quality of Service

Quality of Service (QoS) handling plays an important role in IMT-Advanced network. Chapters 6 and 12 detailed the procedures described by standardization bodies, respectively IEEE and 3GPP, for the QoS handling and bandwidth reservation signaling. Standards, however, do not specify how vendors and operators should implement the schedulers that oversee the prioritization of the different device and user flows, in addition to regulating both access and interference levels in the network. Naturally, such void instigates many researchers to design schedulers that enable IMT-Advanced networks to fulfill their operational objectives. And while this book is aimed at describing the functionalities of the IMT-Advanced technologies, the crucial role of scheduling mandated a dedicated treatment.

The chapter is organized as follows. Section 17.1 discusses scheduling in WiMAX networks, and provides an overview and evaluation of the different proposal that have been suggested for its scheduling. Section 17.2 provides takes on a similar view of scheduling in LTE and LTE-Advanced. Given that the 3GPP technologies utilize different multi-carrier access techniques, OFDMA for the uplink and SC-FDMA, the section discusses scheduling for each connection direction together with its relevant requirements. Section 17.3 provides a further view of evaluations that have been towards comparing the performance of LTE/-A and WiMax in terms of VoIP scheduling and power consumption, in addition a comparison between OFDMA and SCF-FDMA.

17.1 Scheduling in WiMAX

Packet scheduling is the process of resolving contention for shared resources in a network. The process involves allocating resources among the users and determining their transmission order. Scheduling algorithms for a particular network need to be selected based on the type of users in the network and their QoS requirements. For real-time applications such as video conferencing, voice chat,

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

and audio/video streaming, delay and delay jitter are the most important QoS requirements. Delay jitter is the inter-packet arrival time at the receiver and is required to be reasonably stable by real-time applications. On the other hand, for non-real time application such as FTP, throughput is the most important QoS requirement. Some applications, such as web-browsing and email do not have any QoS requirements. In a network, different types of applications with diverse QoS requirements can co-exist. The task of a scheduling algorithm in such a network is to categorize the users into one of the pre-defined classes. Each user is assigned a priority, taking into account his QoS requirements. Subsequently, resources are allocated according to the priority of the users while fairness is observed.

Besides having a very close coupling with the QoS requirements of the users, the design of a scheduling algorithm also depends on the type of the network it is intended for. A wireless network can be categorized into a single-hop or a multi-hop network. A single-hop network contains a central entity such as a BS that makes and delivers decisions to all SSs in its cell. On the other hand, in a cellular multi-hop network, some SSs are not in direct contact with the BS, that is, an object such as a building could be blocking the path from the BS to the SS. In such a network, a RS is used to relay the information to and from these SSs.

Packet scheduling algorithms can usually be distinguished based on their characteristics. In the sequel, we shall review some of the desirable qualities a scheduling algorithm should possess. We also highlight the issues that need to be addressed by these algorithms.

- *Flexibility*: A scheduling algorithm should be able to accommodate users with diverse QoS requirements and also meet the minimum requirements of the users. Ideally, the design of a scheduling algorithm should be flexible enough so that it requires minimal changes to be deployed in a different network or even a different technology.
- *Simplicity*: A scheduling algorithm should be simple, both conceptually and mechanically. Conceptual simplicity allows manageable analysis of the algorithm such that distribution or worst case analysis bound for parameters such as delay and throughput can be derived. Mechanical simplicity allows efficient implementation of the algorithm at a large scale.
- *Protection*: A scheduling algorithm needs to be able to protect well behaving users from sources of variability such as BE traffic, misbehaving users and fluctuations in the network load. Upon admission into the network, users specify a contract they will adhere to, for example, a user will specify the peak rate at which it will send traffic into the network. Sometimes a user will not abide by the contract causing unpredicted fluctuations in the network. A scheduling algorithm needs to ensure that such fluctuations do not affect well behaving users in the network.
- *Fairness*: Besides satisfying the QoS requirements of the users, a scheduling algorithm needs to ensure a reasonable level of fairness is maintained between the users. Fairness measures the difference between the users with respect to the resources allocated to them. In a wireless network, due to the presence of

variations in channel quality, users experiencing poor channel quality might be denied service by the scheduling algorithm. This is because resources allocated to users with inferior channel quality will essentially be wasted as the data will be lost or corrupted prior to reaching the destination. A scheduling algorithm needs to have a mechanism to compensate users that have lost service and maintain fairness between all the users.

- *Link Utilization*: A scheduling algorithm is required to assign bandwidth to the users such that maximum link utilization is realized. Link utilization is a critical property for the service providers as it is directly linked to the revenue generated. A scheduling algorithm needs to ensure that resources are not allocated to users that do not have enough data to transmit, thus resulting in wastage of the resources.
- *Power conservation on the mobile device*: Due to limited power available on the MS, a scheduling algorithm needs to ensure that limited processing is done on this device.
- *Device mobility*: Different cells can have different notions of time, that is, the BSs of different cells are not required to be synchronized. When a MS moves from one cell to another, packets need be time-stamped based on the notion of time in the new cell. Scheduling algorithms that allocate bandwidth to the users according to the time-stamp of the packets (e.g., schedule users based on their packet deadlines) will not function as expected if the packets are not stamped with the correct notion of time.

To evaluate the performance of current WiMAX schedulers, several scheduling algorithms are assessed with respect to the characteristics of the IEEE 802.16 MAC layer and OFDM PHY. The authors of [1] classify the proposals into three categories; homogenous algorithms, hybrid algorithms and opportunistic algorithms. The homogenous and the hybrid categories consist of traditional scheduling algorithms with the hybrid category employing multiple legacy schemes in an attempt to satisfy the QoS requirements of the multi-class traffic in WiMAX networks. The opportunistic category refers to algorithms that exploit variations in channel conditions in WiMAX networks whilst incorporating the QoS requirements in their scheduling design. Representative schemes in each of these categories will be discussed next.

17.1.1 Homogeneous Algorithms

Weighted Round Robin (WRR) and Deficit Round Robin (DRR) algorithms are evaluated in a WiMAX network in [2]. WRR is evaluated for the uplink traffic while DRR is evaluated for the downlink traffic. In WRR, each SS is assigned a weight factor that reflects its relative priority. Priority of the SSs can also be incorporated in the DRR algorithm. DRR allows provision of different quanta for each SS. A higher quantum can be assigned to higher priority SSs. Ruangchaijatupon et al. [3] evaluated the performance of Earliest Deadline First (EDF) algorithm. This is a work conserving algorithm originally proposed for real-time

applications in wide area networks [4]. The algorithm assigns deadline to each packet and allocates bandwidth to the SS that has a queued packet with the earliest deadline. Weighted Fair Queuing (WFQ) is also evaluated and compared with EDF in [4].

Tsai et al. [5] proposed an uplink scheduling algorithm and a token bucket based Call Admission Control (CAC) algorithm. The CAC algorithm assigns thresholds to each class to avoid starvation of lower priority classes. The scheduling algorithm first grants bandwidth to SSs of the UGS class, then allocates bandwidth to SSs of the rtPS class using EDF algorithm and restricting the allocation to the maximum grant size. Finally, the algorithm allocates minimum required bandwidth to SSs of the nrtPS and BE classes.

17.1.2 Hybrid Algorithms

Wongthavarawat and Ganz [6] proposed a hybrid scheduling algorithm that combines EDF, WFQ and FIFO algorithms. The overall allocation of resources is done in a strict priority manner. EDF scheduling algorithm is used for SSs of the rtPS class, WFQ is used for SSs of the nrtPS class and FIFO for SSs of the BE class. Besides the scheduling algorithm, an admission control procedure and a traffic policing mechanism are also proposed.

Vinay et al. [7] proposed a hybrid scheme that uses EDF for SSs of the rtPS class and WFQ for SSs of nrtPS and BE classes. This algorithm differs from [6] in that WFQ is used for SSs of both nrtPS and BE classes and the overall bandwidth is allocated fairly, however, the authors did not describe the mechanism for fair allocations. Settembre et al. [8] propose a hybrid scheduling algorithm that uses WRR and RR algorithms with a strict priority mechanism for overall resource allocation. In the initial stages, resources are allocated on a strict priority basis to SSs of the rtPS and nrtPS classes only. The WRR algorithm is used to allocate bandwidth amongst SSs of rtPS and nrtPS classes until they are satisfied. Any residual bandwidth is distributed between the SSs of the BE class using the RR algorithm.

A vital component of hybrid algorithms is the distribution of bandwidth among the diverse traffic classes. We have selected to evaluate hybrid (EDF + WFQ + FIFO) and hybrid (EDF + WFQ) schemes, which employ different mechanisms of distributing bandwidth among the traffic classes. The hybrid (EDF + WFQ + FIFO) algorithm applies the strict priority mechanism, whereas the hybrid (EDF + WFQ) keeps track of the bandwidth allocated to all service classes and perform dynamic distribution of bandwidth by providing fair service to all traffic classes. In our evaluation, we use the MRTR of a SS as the core of this approach (details were not available in [7]). Specifically, bandwidth is distributed with respect to the relative MRTR of all SSs in a class, that is, the available bandwidth is multiplied by the ratio of sum of MRTR of SSs in a class to the sum of MRTR of all the SSs in the network.

17.1.3 Opportunistic Algorithms

A Cross-Layer scheduling algorithm is proposed in [9] whereby each SS is assigned a priority based on its channel quality and service class. The SS with the highest priority is scheduled for transmission in each frame. The algorithm considers all the required QoS parameters of the scheduling services specified in the IEEE 802.16-2004 standard. Class coefficients are utilized to assign relative priority to the different traffic classes. Rath et al. [10] proposed to use an opportunistic extension of the DRR algorithm with the purpose of satisfying delay requirements of multi-class traffic in WiMAX. The cornerstone of the algorithm is selecting an appropriate polling mechanism. At the beginning of a polling interval, a set of schedulable SSs are selected that constitute a schedulable set. Until the next polling interval, SSs are selected only from this schedulable set.

Niyato and Hossain [11] proposed a joint resource allocation and connection admission control algorithm based on queuing theory. In order to limit the amount of bandwidth allocated per class, a bandwidth threshold is assigned to each class. A utility function is calculated for each SS based on the QoS requirements of the traffic class. Subsequently, bandwidth is allocated based on the utility, giving priority to the SS with the lowest utility.

Singh and Sharma [12] proposed a scheduling algorithm for OFDMA systems with a TDD frame structure for both uplink and downlink traffic in WiMAX. The algorithm allocates bandwidth among the SSs on a priority basis taking into consideration the channel quality, the number of slots allotted to the SS and the total bandwidth demanded by the SS. Kim and Yeom proposed an uplink scheduling algorithm for TCP traffic for the BE class [12]. The proposed algorithm does not require explicit bandwidth request from a SS. It estimates the amount of bandwidth required by the SS based on its current transmission rate. The purpose of the algorithm is to provide reasonable fairness among the SSs based on the min-max fairness criteria while providing high frame utilization.

The Cross-Layer and Queuing Theoretic algorithms provide a good representation of all the schemes in this category. Both algorithms differ with respect to the number of SSs selected for transmission and the QoS parameters incorporated. The queuing theoretic algorithm schedules multiple SSs in each frame whereas the cross-layer algorithm schedules only one SS. The cross-layer algorithm includes both throughput and delay in the priority function of rtPS class but the queuing theoretic algorithm includes only the delay in the utility function of the rtPS class.

The performance of the scheduling algorithms is evaluated under different conditions. These conditions include studying performance of the algorithms under various concentrations of traffic and under characteristics of the IEEE 802.16 MAC layer such as uplink burst preamble, frame length and bandwidth request mechanisms. Table 17.1–17.3 show a summary of the comparison among WiMAX schedulers. A detailed results and discussion can be found in [Pratik].

Table 17.1 Comparison of Homogeneous schemes

	EDF	WFQ	WRR
Intra-class fairness (ertPS/rtPS/nrtPS/BE)	High/High/High/High	Medium/High/Medium/High	High/High/High/High
Inter-class fairness	Low	Medium	High
Frame Utilization	High	High	Low-Medium
Average Throughput (ertPS/rtPS/nrtPS/BE)	High/High/Low/Low	Medium/High/High/Medium	Low/Medium/Medium/Low
Average delay (ertPS/rtPS)	Low/Low	Medium/Low	High/High
Packet loss (ertPS/rtPS)	Low/Low	High/Low	High/Medium

Table 17.2 Comparison of Hybrid schemes

	EDF + WFQ	EDF + WFQ + FIFO
Intra-class fairness (ertPS/rtPS/nrtPS/BE)	High/High/Medium/Low-Medium	High/High/Medium/High
Inter-class fairness	Medium	Low
Frame Utilization	Medium	High
Average Throughput (ertPS/rtPS/nrtPS/BE)	High/Medium/High/Low	High/High/Medium/Low
Average delay (ertPS/rtPS)	High/High	Low/Low
Packet loss (ertPS/rtPS)	Medium/Medium	Low/Low

Table 17.3 Comparison of Opportunistic schemes

	Queuing Theoretic	Cross-Layer
Intra-class fairness (ertPS/rtPS/nrtPS/BE)	Medium/Low-Medium/Low-Medium/High	Low/Low-Medium/Low/High
Inter-class fairness	Medium	Low
Frame Utilization	Low-Medium	Low
Average Throughput (ertPS/rtPS/nrtPS/BE)	High/High/High/Medium	Low/Low/Low/Low
Average delay (ertPS/rtPS)	Medium/Medium	Medium/Medium
Packet loss (ertPS/rtPS)	Medium/Medium	High/High

17.2 Scheduling in LTE and LTE-Advanced

17.2.1 Scheduling the Uplink

LTE uplink scheduler acts as part of the LTE radio resource management functionalities to utilize the available radio resources within the physical uplink shared channel (PUSCH) as efficiently as possible, while satisfying the QoS requirements for active users within the network. Due to the configuration of the radio interface employed in LTE uplink, the uplink scheduler performs user multiplexing in both, time and frequency domains to the available resource blocks (RBs) within 1 TTI. The LTE uplink scheduler designs that have been proposed in literature so far perform scheduling per TTI according to the following phases:

1. *UE Selection*: The first phase of the scheduling operation is to select a subset of the UEs that await transmission to be scheduled for a current TTI. The selection process occurs either by a round robin fashion, a proportional fair criteria, or based on QoS attributes (bit rate, delay, etc.), buffer size, or a combination of these attributes. In a QoS-aware LTE uplink scheduler, the QoS attributes of each scheduler plays an important role in the UE selection process. Since the scheduling process is engaged once every TTI with no consideration for how the bandwidth is to be distributed among the selected UEs, the UE selection can be associated with the Time Domain (TD) scheduling, and be separated from the frequency allocation.
2. *UE-Frequency Multiplexing*: Once the set of UEs to be scheduled are selected, the scheduler distributes the available RBs among the selected UEs. In channel dependent scheduling (CDS), the scheduler exploits the variations of the selective-fading channels to allocate each group of RBs to a UE with the best channel conditions over these RBs. Hence, the multiplexing of UEs over the available radio resources is termed Frequency-Domain (FD) scheduling.

LTE uplink scheduling can be described as a queuing-based operation. The TD-scheduler provides a priority metric to each UE according to a certain criterion. As a result, the UE gets added to a queue based on its assigned priority. The scheduler then selects a subset of UEs with the highest priority from the UE's priority queue. The number of UEs to be selected per TTI depends on the choice of scheduler's implementation. However, it is usually restricted by the resources available in the PDCCH that can be used to communicate resource grants to these UEs simultaneously. The availability of PDCCH resources varies depending on the LTE downlink control channel status and configuration.

For QoS scheduling, TD metrics that are associated with the UEs can be made as QoS-aware metrics, where a per-UE TD metric can be based on either the GBR of the uplink traffic, the delay, or both.

The FD scheduler performs dynamic scheduling to allocate UEs to a portion of the frequency bandwidth based on the uplink channel quality between the UE and the BS. Similar to the TD scheduler, the FD scheduler performs metric

calculations per UE. However, in the case of CDS, the FD scheduler assigns a metric weight for each RB, per UE. The RB metric value depends primarily on the channel condition for the RB of each UE. The FD metric can be based on the CSI, which is the SINR of a RB, or on an estimated achievable throughput of a UE at a specific RB.

The process of calculating both TD and FD metrics for each UE at each RB is defined as the utility function. A utility function is the metric to optimize whatever parameters the LTE system needs to perform at a desired level. The most common objectives that a utility function needs to optimize are spectral efficiency, aggregated throughput, fairness, and QoS guarantee.

The authors of [13] evaluated and categorized the scheduling algorithms currently proposed in literature into three categories based on the method of RB allocation. The first group of scheduling algorithms performs RB allocation with fairly-equal-sized RB groupings. The scheduler divides the available RBs into contiguous chunks, where each chunk represents a contiguous RBs group. Each Resource Chunk (RC) has almost the same number of RBs, where the total number of RCs is set to be the number of UEs. In case the number of RBs is less than the number of UEs, then each RC is set to have only 1 RB. Once RCs are created, the scheduler assigns each RC a metric that is based on an aggregation method chosen by the scheduler (e.g., by summing the RB metrics within the RC, or finding their average).

The authors of [13] studied the performance evaluation of round robin algorithm and a maximum SNR (MAXSNR) [14]. The difference between the two is that latter effectively considers the channel condition within the scheduling decision.

The second group of algorithms performs RB allocation using first maximum search algorithms. Such algorithms perform scheduling according to the following generalized steps:

1. Find the UE-RB with the maximum metric.
2. Allocate the RB to corresponding UE.
3. Expand on currently allocated RB to adjacent RBs for current UE, until an RB is found whose maximum metric belongs to another UE.
4. Allocate the current RB to the new UE if it does not violate the contiguity of resource allocation; otherwise, allocate it to the previous UE.
5. If all UEs are allocated RBs, and there are still RBs left unassigned, allocate the unassigned RBs to the same UEs based on contiguity.

The scheduling algorithms chosen to represent this group are the Greedy algorithm [15], Heuristic Localized Gradient algorithm [16] and First Maximum Expansion (FME) algorithm [17] and its two extensions, Modified FME (M-FME) [17] and Recursive maximum Expansion (RME) algorithm [17]. FME performs RBs allocation using first maximum search algorithm to choose one UE with the maximum utility, while extension of FME chooses the first two UEs with the maximum utility. The RME removes UEs, which are already allocated

RBs that have maximum metric associated with these UEs and runs a recursive of the maximum metric on the remaining RBs.

The third group of scheduling algorithms is the global-metric driven algorithms. Rather than starting with the UE-RBs with the maximum algorithm, the algorithms find the UE-RB allocations provide a maximum global metric. The algorithms find a combination of possible allocations, and select the one with maximum global metric. For example, a UE-RB with the highest sum of metrics of UE-RB pairs among the examined combinations.

The algorithms of this category either work on individual RBs, or assign them in RC. Unlike the schedulers of the first group, the globally driven algorithms can either create RCs to be fairly equal in sizes, or RCs sizes that are independently set based on some system criteria. The following steps describe the general steps taken by such algorithms:

1. For each RB or RC, find the UEs with the first maximum metrics.
2. Determine the possible combinations of resource allocations using the selected UEs for each RB, or RC, and construct a search tree.
3. Select the optimal allocation pattern by finding the search tree branch with the maximum, or minimum, weight such that it maximizes the utility function.

Proportional Fair Binary Search Tree (PF-BST) algorithm [15] and Minimum Area Difference (MAD) algorithm [17] are the two algorithms evaluated as an example of this group.

The performance evaluation measures used in [13] are the throughput and the spectral efficiency. The performance evaluation results show that the above mentioned algorithms exhibit comparable performance irrespective of their category. Given the fact that, the RME algorithm entails comparable performance to the other algorithms but distinguished by the least complexity, the authors concluded that the RMS performance is promising as it indicates that acceptable performance can be achieved at low processing requirements.

17.2.2 Scheduling the Downlink

The work presented in [18] investigated multiple packet scheduling algorithms originally proposed for single carrier downlink transmission and good candidates for use in LTE. The authors studied the usability of these algorithms for the implementation of downlink LTE transmission. The algorithms are selected with the aim of maximizing throughput along with fairness. The algorithm studied is the maximum rate (Max-Rate) [19] algorithm, which priorities users with the highest reported instantaneous downlink SNR values. This algorithm maximizes the network throughput. However, it results with low fairness performance. The RR [20] algorithm is chosen to study the performance of LTE scheduler if the fairness is the main objective to achieve. RR is simple in implementation and provides for fairness by allocating an equal share of transmission times to each user. It is obvious that the RR algorithm while meeting the fairness measures it performs poorly

when maximizing the throughput is the objective. The authors of [18] studied the performance of Proportional Fair (PF) [21] algorithm to provide a balance between throughput and fairness. The proportional fairness algorithm keeps track of the average data rate of the user over a predefined window. Users are prioritized based on the ratio of their instantaneous attainable data rate to their average data rate in attempt to maximize throughput along with fairness. The above algorithms do not count for the delay requirements of users. Hence, the authors included the Maximum-Largest Weighted Delay First (M-LWDF) [22] in the study to investigate the support of real time applications with delay constraints. The algorithm incorporates the head of line packet delay with the PF mechanism discussed above to prioritize users, hence, strike a balance between low packet loss, fairness and throughput. The exponential/proportional fair (EXP/PF) [23, 24] algorithm schedule real and nor-real time applications users. Real time users receive a higher priority than non-real time users when their head-of-line packet delays are approaching the delay deadline. The simulation environment consists of one cell and 80–120 users constantly moving at speeds between 1- 100 km/h in random directions. The performance evaluation is carried out with one application under investigation in the network, video streaming. The authors claim based on their simulation results that the M-LWDF algorithm outperforms other packet scheduling algorithms by providing better fairness and higher throughput which allows accommodating larger number of users.

17.3 Quantitative Comparison between LTE and WiMAX

In this section we present the research attempts for comparing the performance of some LTE and WiMAX functionalities in the same experimental setup. We present works addressed the performance evaluation of the VoIP scheduling in WiMAX and LTE, the power consumption of the WiMAX and LTE BS and the access methods used in both technologies, the OFDMA in WiMAX and the SC-FDMA in LTE.

17.3.1 VoIP Scheduling in LTE and WiMAX

The work in [25] studies the performance of VoIP scheduling for TDD-LTE and IEEE 802.16m. The work compares the two technologies performance schedulers in serving VoIP applications. The semi-persistent schedulers are only defined by LTE. They are useful in serving the delay-intolerant VoIP traffic, since the semi-persistent schedulers reduce the amount of control signaling while maintaining an acceptable level of output quality. The study implements the semi-persistent schedulers for VoIP packets initial transmission and a dynamic scheduler for serving the VoIP packets retransmissions. The authors concluded that the IEEE802.16m persistent scheduler has higher capacity than that of LTE-TDD in the uplink, because LTE-TDD implements SC-FDMA RB allocation in the uplink, hence, LTE-TDD is required to implement sort of

inter-cell interference coordination algorithm to mitigate interference. However, 802.16m allocates RB using OFDMA. In dynamic scheduling in both uplink and downlink, the authors observed that LTE-TDD outperforms 802.16m, because LTE can effectively achieve better frequency selectivity gain over 802.16m because the MCS in LTE are finer than those in 802.16m.

17.3.2 Power Consumption in LTE and WiMAX Base Stations

The authors of [26] studied the power consumption of outdoor LTE and WiMAX BSs. The authors observed that a WiMAX BS is more energy efficient than an LTE station. LTE BS has power consumption higher by 29 % than WiMAX BS with a range-coverage 27 % lower than WiMAX for a 1 × 1 SISO setup. The power consumption increases for LTE in a 4 × 4 MIMO setup to 30–32 % for an increase of coverage of 132 %, while for the same coverage increase, the power consumption of a WiMAX station increases with only 8 %.

17.3.3 Comparing OFDMA and SC-FDMA

The work in reference [27] reported the performance evaluation between two multiple access techniques used in LTE and WiMAX; OFDMA used in uplink and downlink transmission in WiMAX and SC-FDMA used uplink transmission in LTE. The results of the performance evaluation do not prefer a technique over the other, neither of the two technologies has better performance all the time over the other. For example, OFDMA has better performance with high-order modulations. Meanwhile SC-FDMA has better performance with low-order modulation specifically QPSK. Hence, OFDMA can offer higher cell throughput, while SC-FDMA can provide larger cell coverage.

References

[1] Najah Abu Ali, Pratik Dhrona and Hossam Hassanein, " A performance study of uplink scheduling algorithms in point-to-multipoint WiMAX networks", *Computer Communications*, Volume 32, Issue 3, Adaptive Multicarrier Communications and Networks, 25 February 2009, pp. 511–21.

[2] C. Cicconetti, A. Erta, L. Lenzini and E. Mingozzi, "Performance Evaluation of the IEEE 802.16 MAC for QoS Support", *IEEE Transactions on Mobile Computing*, vol. 6, no.1, pp. 26–38, January 2007.

[3] N. Ruangchaijatupon, L. Wang and Y. Ji, "A Study on the Performance of Scheduling Schemes for Broadband Wireless Access Networks", *Proceedings of International Symposium on Communications and Information Technology*, pp. 1008–12, October 2006.

[4] D. Ferrari and D. Verma, "A scheme for real-time channel establishment in wide-area networks", *IEEE Journal on Selected Areas in Communications*, vol. 8, no.3, pp. 368–79, April 1990.

[5] T. Tsai, C. Jiang and C. Wang, "CAC and Packet Scheduling Using Token Bucket for IEEE 802.16 Networks", *Journal of Communications*, vol. 1, no. 2., pp. 30–7, May 2006.

[6] K. Wongthavarawat, and A. Ganz, "Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems", *International Journal of Communication Systems*, vol. 16, issue 1, pp. 81–96, February 2003.

[7] K. Vinay, N. Sreenivasulu, D. Jayaram and D. Das, "Performance evaluation of end-to-end delay by hybrid scheduling algorithm for QoS in IEEE 802.16 network", *Proceedings of International Conference on Wireless and Optical Communication Networks*, 5 pp., April 2006.

[8] M. Settembre, M. Puleri, S. Garritano, P. Testa, R. Albanese, M. Mancini and V. Lo Curto, "Performance analysis of an efficient packet-based IEEE 802.16 MAC supporting adaptive modulation and coding", *Proceedings of International Symposium on Computer Networks*, pp. 11–16, June 2006.

[9] Q. Liu, X. Wang and G. Giannakis, "Cross-layer scheduler design with QoS support for wireless access networks", *Proceedings of International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks*, 8 pp., August 2005.

[10] H. Rath, A. Bhorkar and V. Sharma, "An Opportunistic uplink Scheduling Scheme to Achieve Bandwidth Fairness and Delay for Multiclass Traffic in Wi-Max (IEEE 802.16) Broadband Wireless Networks", *Proceedings of IEEE Global Telecommunications Conference*, pp. 1–5, November 2006.

[11] D. Niyato and E. Hossain, "A Queuing-Theoretic Optimization-Based Model for Radio Resource Management in IEEE 802.16 Broadband Wireless Networks", *IEEE Transactions on Computers*, vol. 55, no. 11, pp. 1473–88, November 2006.

[12] V. Singh and V. Sharma, "Efficient and fair scheduling of uplink and downlink in IEEE 802.16 OFDMA networks", *Proceedings of IEEE Wireless Communications and Networking Conference*, pp. 984–990, September 2006.

[13] K. Elgazzar, M. Salah, A.M. Taha and H. Hassanein, "Comparing uplink schedulers for LTE". In *Proceedings of the 6th international Wireless Communications and Mobile Computing Conference*, Caen, France, pp. 189–93, 2010.

[14] F. Calabrese, P. Michaelsen, C. Rosa, M. Anas, C. Castellanos, D. Villa, K. Pedersen, and P. Mogensen, "Search-tree based uplink channel aware packet scheduling for utran lte," in *Vehicular Technology Conference, 2008*. VTC Spring 2008. IEEE, pp. 194953, 11–14, 2008.

[15] S.-B. Lee, I. Pefkianakis, A. Meyerson, S. Xu, and S. Lu, "Proportional fair frequencydomain packet scheduling for 3gpp lte uplink," in INFOCOM 2009, IEEE, pp. 2611–15, 19–25 2009.

[16] L. Ruiz de Temino, G. Berardinelli, S. Frattasi, and P. Mogensen, "Channel-aware scheduling algorithms for sc-fdma in lte uplink," in *Personal, Indoor and Mobile Radio Communications, 2008*. PIMRC 2008. IEEE 19th International Symposium on, pp. 1–6, 15–18 2008.

[17] M. Al-Rawi, R. Jantti, J. Torsner, and M. Sagfors, "Opportunistic uplilnk scheduling for 3g lte systems," in *Innovations in Information Technology, 2007*. IIT '07. 4th International Conference on, pp. 705–9, 18–20 2007.

[18] H.A.M. Ramli, R. Basukala, K. Sandrasegaran and R. Patachaianand, "Performance of well known packet scheduling algorithms in the downlink 3GPP LTE system," *IEEE 9th Malaysia International Conference on Communications (MICC), 2009*, pp. 815–20, 15–17 Dec. 2009.

[19] B.S. Tsybakov, "File Transmission over Wireless Fast Fading downlink," *IEEE Transactions on Information Theory*, vol. 48, pp. 2323–37, 2002.

[20] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G Evolution: HSPA and LTE for Mobile Broadband*, 1st edn: Elsevier Ltd., 2007.

[21] A. Jalali, R. Padovani, and R. Pankaj, "Data Throughput of CDMA HDR a High Efficiency-High Data Rate Personal Communication Wireless System," *in IEEE 51st Vehicular Technology Conference Proceedings*, Tokyo, 2000, pp. 1854–8.

[22] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing Quality of Service over a Shared Wireless Link," *IEEE Communications Magazine*, vol. 39, pp. 150–4, Feb. 2001.

[23] J.-H. Rhee, J. M. Holtzman and D.K. Kim, "Performance Analysis of the Adaptive EXP/PF Channel Scheduler in an AMC/TDM System," *IEEE Communications Letters*, vol. 8, pp. 4978–80, Aug. 2004.

[24] J.-H. Rhee, J. M. Holtzman, and D. K. Kim, "Scheduling of Real/Non-real Time Services: Adaptive EXP/PF Algorithm," in *The 57th IEEE Semiannual Vehicular Technology Conference*, vol. 1, 2003, pp. 462–6.

[25] Zhijie Wang, Yafeng Wang and Fei Wang, "Comparison of VoIP capacity between 3G-LTE and IEEE 802.16m," *2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 2192–6, 13–16 Sept. 2009.

[26] M. Deruyck, W. Vereecken, E. Tanghe, W. Joseph, M. Pickavet, L. Martens and P. Demeester, "Comparison of power consumption of mobile WiMAX, HSPA and LTE access networks," *2010 9th Conference on Telecommunications Internet and Media Techno Economics (CTTE)*, pp. 1–7, 7–9 June 2010.

[27] C. Ciochina and H. Sari, "A review of OFDMA and single-carrier FDMA," *2010 European Wireless Conference (EW)*, pp. 706–10, 12–15 April 2010.

18

The Market View

In October 2010, the ITU-R recognized 3GPP's LTE-Advanced and IEEE's WirelessMAN-Advanced interfaced as IMT-Advanced technologies. At the same time, the ITU-R recognized this qualifies both interfaces to be "true 4G" technologies [1].

Operators are currently dealing with a strong demand for broadband Internet, one that is evident across the different sectors. In 2010, demand for mobile data more than doubled, and underwent an overall growth rate of 190 %. Video streaming, in particular, assumed a substantial portion of this increased demand, bearing on almost 40 % of all application traffic. [2] By the end of the same year, the ITU Development (ITU-D) sector estimates a total of 5.3 billion mobile cellular subscriptions to have been made, with 940 million subscriptions being made to 3G networks with access available to around 90 % of the world population [3]. Such indicators are motivating operators to deploy more and more 3G networks to respond to this increased demand.

At the same time, and as the ITU-D equally notes, the year 2010 marks the end of "double-digit mobile growth rate" overall, with the market reaching saturation in developed countries (an average of 116 subscriptions per 100 inhabitants), but with reasonable growth still to be expected in Asia and Pacific countries.

In what follows, we offer a detailed overview of the market status and outlook for IMT-Advanced technologies. In the next section, we discuss the status of the mobile and wireless market today, in terms of both networks and handsets, and considering other relevant aspects. We also draw on the possible evolution tracks that will be taken by operators in getting to 4G networks. The chapter is organized in two parts. The first part, Section 18.1 the recent and the near future efforts in the market as they move towards deploying IMT-Advanced networks. Especially, it elaborates on the pre-IMT-Advanced activity, and the possible evolution tracks that operators might take towards next genet generation networks. The second part, Section 18.2, builds on the market description provided in the first section

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

and describes the outlook for IMT-Advanced markets. It also describes the interplay and the effect of spectrum deregulation, the proliferation of "small cells", including relay stations, femtocells and WiFi Spread. The section also touches on operator readiness 4G, and important roles backhaul investments and patent management will play in accelerating the deployment of IMT-Advanced networks. Finally, the section concludes with a brief comment on the road ahead.

18.1 Towards 4G Networks

Until very recently, the bulk of the international infrastructure for cellular networks was based on 2G networks. However, strong migrations to 3G networks occurred and are still underway, to CDMA based networks (3GPP and 3GPP2) and their respective evolutions. 3GPP standards ratified in 2006 and 2007 for High Speed Downlink and Uplink Packet Access (HSDPA and HSUPA), collectively called High Speed Packet Access (HSPA), have been ratified through several 3GPP Releases, with HSPA Evolution (HSPA+) ratified in 2007. Strong deployment activity followed shortly after, for example, US AT&T's HSDPA was launched in 2009, and T-Mobile announcing its HSPA+ in 2010 [4].

There is currently great activity in deploying the latter evolutions of 3G technologies, including both HSPA+ and LTE. The advent of the smart phones, such as Apple's iPhone and Android-based phones from various vendors [5]; in addition to other devices such as pads and 3G enabling USB connectors (dongles), are all accelerating such deployments. A boom in sales for smartphone devices can definitely be observed in the global market. Some earlier estimates had smartphone unit sales to be around 200 million units [4]. However, almost 300 million smartphone units were shipped last year, with overall handset shipments reaching 1.3 billion units. While iPhone assumes a greater market share, Android has undergone great growth in market share in 2010, with the remainder of the market shared by Nokia and Research in Motion (RIM).

The true impact of smartphone devices is what facilitated 3G usability to the general cellular user, employing both attractive and high resolution interfaces, and the notion of downloadable applications serving various functionalities and attending to particular user demands. This niche-driven or micro-trend market offering of mobile applications provided a great flexibility, strongly parting from fixed, pre-determined setups that were previously the norm in mobile handset market. Despite a tremendous growth in previous years, the mobile application business is expected to experience a 190 % growth, surpassing $ 15 billion in 2011 [6].

The GSA reports that, as of January 2011, there are 383 WCDMA networks commercially launched in 156 countries. Of these, there are 380 operators that have launched HSPA and 103 networks with HSPA Evolution (HSPA+). Such growth is reflected in terms of devices, with around 3000 devices launched by 255 supplies – including 92 HSPA+ devices. [7, 8].

With standards for IMT-Advanced networks ratified by both the 3GPP and the IEEE, it is worth understanding the possible evolution tracks of existing cellular infrastructures.

The "pure 3GPP" track starts at the 2G and 2.5 GSM based digital networks including GPRS and EDGE. Operators of such networks migrated to the CDMA-based 3G UMTS networks, for which the HSDPA and HSPA upgrades did not require much overhaul in the network infrastructure. Migration paths from HSPA and HSPA+ to LTE are possible. In other words, an operator as the option of either expanding its 3G investments from HSPA to HSPA+ then LTE, or migrating directly to LTE.

Meanwhile, the "mixed 3GPP/3GPP2" track starts at 3GPP2'S 2G CDMA One network and through its 2.5G cdma2000. 3GPP2's 3G network and its evolution including the 1xEV-DO Revisions 0, A and B have been established. 3GPP discontinued its pre-4G and 4G efforts, and operators of 3GPP2 networks will hence migrate to LTE through either 1xEV-DO Rev A or Rev B.

As for WiMAX, the WiMAX Forum reports 583 network deployments in 150 countries. The report, however, includes both mobile and fixed WiMAX deployments [9]. Most notable of these deployments is Sprint's commitment to build a US-wide WiMAX deployment with a networks built by Clearwire, which has is to be maintained at least through 2012.

The evolution track for mobile WiMAX is quite straightforward, as operators can migrate directly from the WirelessMAN (IEEE 802.16e or 802.16-2009) to the WirelessMAN-Advanced (IEEE 802.16m). Similarly, for LTE, operators can migrate from LTE to LTE-Advanced. With that said, it should be noted that migration to LTE will require extensive infrastructure upgrade, especially in the backhaul network. [10]. Note that cross migrations between WiMAX and LTE have been indicated to be possible, either directly from WirelessMAN (IEEE 802.16e) to TD-LTE, or through A WirelessMAN to WirelessMAN-Advanced to TD-LTE evolution chain [11].

18.2 IMT-Advanced Market Outlook

At the time of writing this book, there remains strong speculation as to the future of IMT-Advanced networks. However, there are several strong indicators that the market has already sided with LTE and LTE-Advanced as the network of choice when it comes to evolving existing cellular infrastructures. For example, the GSA reports pre-commitments (trials) and commitments by 196 operators in 75 countries. GSA also reports, as of March 2011, the launch of almost 100 devices into the market, including 6 smart phones, 7 tables and 22 Modules. For WiMax, a commitment to mobile WiMAX (IEEE 802.16e or 802.16-2009) comprises a commitment to the WirelessMAN-Advanced. Therefore, the above commitments noted by the WiMAX Forum, which include 150 deployments worldwide, stand as commitments to WiMAX's Advanced evolution. Most recently, Sprint has

infused a 1 billion dollar investment into Clearwire (56 % owned by Sprint), giving a strong thrust to Clearwire's WiMAX deployment [12]. Such infusion has muted expectations of Clearwire's support for LTE as it did announce in 2010 LTE trials. Sprint's own success with its WiMAX deployment and its signature EVO smartphone is being viewed as an indicator for the potential success for 4G networks. The status for the supporting WiMAX devices, however, remains unclear.

For a long time, it has been held that WiMAX has a definite and powerful time-to-market advantage over LTE and LTE-Advanced networks. This is because of WiMAX's maturity as a technology, but also due to its higher readiness to be deployed. On the other hand and despite the many commitments, LTE remains at the trial stage. Recent market reports, however, indicate that WiMAX's advantage may be short-lived and that WiMAX, despite prospects of growth, will be eventually eclipsed by LTE's growth, which is projected to take an exponential lead starting by 2012. Market projections include 14.9 million global WiMAX subscribers by the end of 2011, but no more than 50 million subscribers by 2014 [13]. In 2012, however, LTE will take a strong market lead that reaches between 16 to 50 million by year's end [13, 14].

More generally speaking, both technologies share common facilitators and inhibitors when it comes to deployment. The following discusses some of these common aspects, including spectrum allocation, small cell concept, WiFi spread, the backhaul bottleneck, and operator readiness for 4G investments.

18.2.1 Spectrum Allocation

Spectrum allocation and management, for example, have been observed to be an impediment when it comes to deploying LTE networks, particularly in Europe. This is especially the case given recent activating in auctioning spectrums in the 700 to 1000 MHz range, in addition to the relevant auctioning policies and guidelines that have been set by the different regulators. In the summer of 2011, the European Parliament will meet to decide on the fate the 800 MHz, and whether it will be possible to harmonize its allocation for broadband services. At the same time, UK's Ofcom has decided to cap spectrum purchases in the upcoming auction for LTE spectrum, and fears from monopoly in the upcoming French auction have raised requests for a similar policy. [15] A highly relevant debate that is currently taking place is one that is contemplating new models of spectrum allocations, management and trading.

The interest in sub 1000 MHz spectrum bands stem from the hope to reduce deployment costs. At higher frequencies, signals attenuate much faster, with indoor performance suffering the most. At low frequencies, however, a lower number of base stations is required to cover the same the area. Areas that have been underserved until now, such as rural and suburbia, would therefore benefit greatly from such low frequency allocations. This tradeoff between frequency and deployment costs, however, should be viewed while minding capacity. As will be noted on the next page, high frequency and short range coverage can

be especially effective in providing high capacity wireless links, especially when advanced antenna techniques, that is, MIMO, are exploited [16].

18.2.2 Small Cells

On the facilitation side, the "small cell" phenomena seem to be gaining great popularity on both the operators' and the users' side. Both 3GPP and IEEE have made extensive support for accommodating "small cells" that can be deployed either as femtocells, relay stations or out-of-band WiFi cells to which the IMT-Advanced users can be migrated to. Measures for inter-technology handovers, mandated by IMT-Advanced requirements, means that users can migrate their active connections to WiFi networks. Small cells benefit from the above noted advantages, achieving high capacity gains by both reduced coverage and limited subscriber access. In the case of WiFi networks, great cost savings are made as WiFi nodes operate in the unlicensed ISM band.

In addition to their other advantages, the economic advantages of relay stations have been repeatedly demonstrated in various technoeconomic evaluations, and for both transparent and non-transparent relaying [17]. Such advantages have been demonstrated through different evaluation scenarios, for example, rural, suburban and urban, and under different antenna structures. It was found, for example, that relay stations can provide substantial gains in rural deployments made under high frequency spectrum allocations. Such deployments would naturally use non-transparent deployments as the interest would be largely in expanding coverage. Meanwhile, relay stations (mostly transparent) combined with the advanced antenna technique prove more useful in denser deployments commonly made in suburban and urban areas [18].

Offloading to femtocells and WiFi will reduce the traffic load on an operator's backhaul network. In the various femtocell offerings that have been made in the market, operators may also gain increased revenues from fetmocells through monthly fees, greater loyal and reduced churn [4]. Certain studies, however, have cautioned from generalizing cost reductions in all deployment scenarios. For example, it is possible to such gains in areas where there is a sparse macrocell deployment. The deployment of femtocells in these scenarios would overcome the indoor coverage challenge, and offer enhanced service rates – all at a much a reduced cost than increased macrocell deployments. Meanwhile, in areas where there is already a reasonable macrocell density, the benefit of femtocell deployment may be marginal [19].

18.2.3 The WiFi Spread

Meanwhile, WiFi deployments are continuing a steadfast wide deployments and at an international scale. It is now common to expect free or low-cost WiFi access in nearly all possible venues. The technology's low deployment cost, in additional to minimal requirements of operational intervention, enable WiFi

access providers to deploy very large networks in short durations. Vendors such as BelAir Networks, for example, continue to grow in their market share with their focus on small cells including both WiFi and femtocells, but largely the former. For example, BelAir offers Plug n' play WiFi routers and femtocells with Power Line Communications that can be fit in very short times on existing power, diminishing infrastructure and rental costs for operators. Meanwhile, WiFi access providers such as Boingo, continue to expand their own and partner networks – negotiating over 125 000 locations around the word as of early 2010. WiFi deployments also continue to gain strong grounds in the enterprise, with estimates for in-building wireless installations expected more than $ 6 billion in 2010 [4].

Such large-scale deployments of WiFi networks open the possibilities of new business models. Through joint resource managements, operators deploying mixed access technologies will be able to manage the resources of the technologies in a joint manner, migrating users from one technology to the other based on operational objectives. At the same time, companies are now offering services that facilitate smooth inter-technology handovers, exploiting the recent advances and standardizations that have been made available [20].

18.2.4 The Backhaul Bottleneck

One notable impediment to the realization of the full capacities of both IMT-Advanced technologies is the incapability of operator's backhaul networks to cope with the advances at the radio interfaces. Observations that have been made from the onset of the race towards the IMT-Advanced standardization still stand true today – that in terms of capabilities, both IMT-Advanced technologies stand on equal footing in terms of general performance and compliance to the overall ITU-R requirements. However, as is noted [10], many carriers are constrained by 1.5 Mbps (T1) backhaul, creating a definite limit on network performance, regardless of capabilities of the chosen radio interface. Existing technologies, including both packet microwave and fiber optics, are more than capable of answering the project user demands for IMT-Advanced. Investments decisions, however, have mostly been delayed by the debate on the technical qualities of both access technologies [21].

18.2.5 Readiness for 4G

A note should be made here on the reluctance of incumbent cellular operators to invest in new infrastructures. Substantial investments were made by these operators in 3G networks, both backhaul and infrastructures, which partly justifies the delay as the full revenue potential of these networks is yet to be realized. It is hence that many operators around the world have sought government support, especially through the most recent economic downturn. In many countries, including the US, Canada, Europe, Australia and New Zealand, economic surplus

packages have dedicated funds to support expanded broadband infrastructures, especially when it came to rural and sub urban areas. Such funds, in addition to dedicated partnerships between governmental sectors at the different levels (i.e., federal, provincial, and municipal) and the private sector, are accelerating greater access to broadband Internet in many areas. At the same time, such initiatives are somewhat lessening the burden of expanding operator infrastructures.

Another factor that impeded the realization of profits in deploying 3G networks stems from issues in patent management [22]. It is hence that calls were made to "pool" the patents for LTE, and several initiatives – including the IIT initiative in Canada, where made with this objective. Patent pooling enables several companies to utilize each other's patents when producing a certain product, substantially reducing the royalty fees.

18.3 The Road Ahead

The two IMT-Advanced technologies share many characteristics, as noted throughout this book. It has also been noted how little are the differences that exist between them. While much of the cellular market have chosen the 3GPP evolution track for future cellular networks, this does not mean the end of WiMAX as a technology. WiMAX applicability for fixed wireless broadband, in addition to its attractiveness to "Greenfield" wireless operators in various settings, indicates a sustainable existence for the IEEE technology. The general expectation, therefore, is that of co-existence, where each technology is appropriately deployed to achieve certain objectives – possibly non-overlapping.

References

[1] ITU Newsroom, http://www.itu.int/net/pressoffice/press_releases/2010/40.aspx.
[2] Allot Communications, "Allot MobileTrends: Global Mobile Broadband Traffic Report", H2, 2010. (http://www.allot.com/MobileTrends_Report_H2_2010.html).
[3] ITU-D, "The World in 2010: ICT Facts and Figures," October, 2010 (http://www.itu.int/ITU-D/ict/material/FactsFigures2010.pdf).
[4] Plunkett's Wireless, Wi-Fi, RFID & Cellular Industry Almanac 2011, Plunkett Research, Ltd.
[5] http://www.android.com/.
[6] Gartner Research, http://www.gartner.com/it/page.jsp?id=1529214.
[7] Fast Facts, available at GSA Statistics http://www.gsacom.com/news/statistics.php4.
[8] 3G/WCMA commercial deployments, available at GSA Statistics http://www.gsacom.com/news/statistics.php4.
[9] WiMAX Forum, Industry Research Report, March 2011, available at http://www.wimaxforum.org/resources/research-archive.
[10] G. Lawton, "4G: Engineering versus Marketing," *IEEE Computer Magazine*, Volume 44, Issue 3, pp. 14–16, March 2011.
[11] Aviat, "WiMAX 16e: Evolutionary Choices between 16m and TD-LTE", Whitepaper, July 2010. (http://www.portals.aviatnetworks.com/exLink.asp?9023976OP36K25I69957576).
[12] http://www.gottabemobile.com/2011/04/20/sprints-1-billion-infusion-in-clearwire-demonstrates-wimax-commitment/.
[13] iSuppli Market Research, http://www.isuppli.com/Mobile-and-Wireless-Communications/News/Pages/LTE-to-Overcome-WiMAX-and-Dominate-4G-Shipments.aspx.

[14] ABI Research, http://www.abiresearch.com/press/3672-16+Million+Mobile+LTE+Subscribers+by+Year%92s+End.
[15] http://www.abiresearch.com/press/3672-16+Million+Mobile+LTE+Subscribers+by+Year%92s+End.
[16] G. Goth, "Something's in the Air: Broadband Advances Depend on Wireless," *IEEE Internet Computing Magazine*, Volume 14, Issue 5, pp. 7–9, September 2010.
[17] Y. Yang et al, "Relay Technologies for WiMAX and LTE-Advanced Mobile Systems," *IEEE Communications Magazine*, Volume 47, Issue 10, pp. 100–5, October 2009.
[18] A. Moral et al., "Technoeconomic Evaluation of Cooperative Relaying Transmission Techniques in OFDM Cellular Networks," *EURASIP Journal on Advances in Signal Processing*, Volume 2011, Article ID 507035, 23 pages, 2011.
[19] J. Markendahl and O. Makitalo, "A Comparative Study of Deployment Options, Capacity and Cost Structure for Macrocellular and Femtocell Networks," *in Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 145–50, September 2010.
[20] D.E. Charilas and A.D. Panagopoulos, "Network Selection Problem: Multiaccess Radio Network Environments," *IEEE Vehicular Technology Magazine*, Volume 5, Issue 4, pp. 40–9, December 2010.
[21] http://next-generation-communications.tmcnet.com/topics/nextgen-voice/articles/116462-wimax-vs-lte-does-it-matter.htm.
[22] Z. Abichar, J.M. Chang and C-Y. Hsu, "Wimax vs. LTE: Who Will Lead the Broadband Mobile Internet," *IEEE IT Professional Magazine*, Volume 12, Issue 3, pp. 26–32, May–June 2010.

19

The Road Ahead

The contrast in capabilities between IMT-Advanced and its predecessor is remarkably exciting. Results from the various Evaluation Groups reporting to the ITU-R WP 5 continue to indicate that the two technologies more than satisfy the ITU-R requirements [1], which were designed to address the increasing demand for mobile traffic. To put the demand expectations into perspective, it is helpful to note that a 26-fold increase in mobile data traffic is expected by 2015, reaching a rate of 6.3 exabytes[1] per month [2]. With the world population estimated to grow to a range of 7.2 to 7.5 Billion people [3], an estimate is made there will be as much as 1 mobile unit or device per capita connecting wirelessly. Estimates are also predicting that 1.3GB per month generated per smartphone, with video taking up to two thirds of the traffic. In 2020, more than 50 Billion devices will be connected to the Internet and serving a population in the range of 7.5 to 7.9 Billion – almost six devices per capita [4].

Towards this vision, the earlier deployments of IMT-Advanced would have been made, with great advancements made in both the wireless and the wired Internet. The deployments will provide substantial understanding and experience of how OFDMA operates in practice, which is currently lacking. The use of heterogeneous access networks is also projected to be the norm, with different access technologies aimed at different connection requirements. In the meantime, policies and technologies currently investigated for combating spectrum will slowly emerge, and initial large scale realizations of the adaptive and opportunistic software-defined or cognitive radios will be made. Together with these physical layer advances, especially in cooperative MIMO communication, dynamic spectrum access and allocation will open the door for higher capacity communication. It is these capacities that will make possible high bandwidth transmissions, both in the downlink and the uplink, in addition to supporting the transport of massive amounts of information. At the radio access interworking and

[1] An Exabyte is 2^{60} bytes.

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.

backhaul level, advances will facilitate a more capable network-end management of network functionalities that is fitting to the multitude of devices to be communicating through the network.

At a larger scale, earlier forms of network intelligence will appear that will facilitate much desirable characteristics of network autonomy. Such characteristics include the currently deliberated aspects of self-optimization and self healing. This autonomy will depend on processing massive amount of information that will be already traversing the network, generated either by passive sensing or through active Machine-to-Machine (M2M) communications. Many of the recently starting initiatives will ensures that such processing is made in a manner that preserves the integrity and privacy of the processed information, while achieving the desired network performance and user satisfaction objectives. At the same time, operators and vendors will begin employing mechanisms for reducing energy requirements, both per-unit and for networks as a whole. Such "greener operation", however, will not be at the cost of network reliability.

The following describes the enablers of this vision, together with the challenges faced to realize its practicality. The chapter is organized as follows. Section 19.1 discusses how IMT-Advanced network will realize reliable network capacity that fulfills the increasing user demands in a cost-efficient manner. Section 19.2 then elaborates on the access heterogeneity, and how exploiting the availability of multiple access technologies will materialize over the next ten years, especially as smarter, multiple-mode devices are introduces. The role played by cognitive radios, and impact of dynamic spectrum allocation and access will be highlighted in Section 19.3, while aspects and applications of in-network intelligence will be discussed in Section 19.4. Advances in network access infrastructure, and the importance of ever "flattening" network architectures are all discussed in Section 19.5. Meanwhile, the complexity of resource allocation, and the efforts made to combat them are discussed in Section 19.6. Finally, Section 19.7 discusses how more and more elements of IMT-Advanced networks will be "green". The section also notes the basic tradeoffs that ground the expectations for how green IMT-Advanced networks will be.

19.1 Network Capacity

Chapter 2 discussed enabling technologies and advances that were adopted for LTE, WiMAX and their IMT-Advanced successors. As noted before, several technologies were sought in order to enhance the capacity of access networks at a cost efficient manner. Without doubt, the choice of multi-carrier access techniques will offer both great flexibility and reliability in such direction. However, it is in advanced antenna and network configurations that substantial capacity gains are achieved. Already, the notion of small cells – through the in-band femtocells or out-of-band WiFi networks – are already beginning to play an important role in today's networks. The importance of small cells in the next few years can be highlighted by estimates of the amount of data they are expected to support −800 million terabytes per month by 2015 [2].

Other advances will come at higher costs, including the use of relaying techniques and cooperative MIMO. As noted in Chapter 2, it is generally understood that such "meshed" wireless communications can provide substantial gains. Relaying, for example, combats path loss and shadowing loss through the breaking down of the wireless link into smaller and reliable segments. Similarly with MIMO, which have shown great versatility in either mitigating interference or enhancing the reliability of the wireless link. And while for some of these advances the limits on possible gains are yet to be figured [5], the practicality of achieving these gains will be slowly evaluated over the next ten years as they are introduced to actual deployments. Certain issues, such as finding deployable mechanisms for resource allocations, remain unresolved. More critically, it will be important to demonstrate that capacity gains made exhibit reliability and cost efficiency.

19.2 Access Heterogeneity

LTE and LTE-Advanced are complemented by an IP-based network core, the EPC. There is also strong IP-based internetworking in WiMAX. Such support will be crucial in creating heterogeneous network composites – not only for user access, but for generalized device access. As noted in Chapter 16, work within the 3GPP and 3GPP2, in addition to the efforts in IEEE 802.21 or Media Independent Handover, are all aimed at supporting inter-technology handovers at the access level. There are also efforts including those of the IEEE P1900 working group that are aimed at, among other things, enhancing operational coexistence between the different radio technologies.

A definite trend that is to grow over the coming years is the addition of satellite networks to the existing heterogeneity. Traditionally, and despite their great bandwidths, satellites have been avoided for user- and device level access due to both their cost and delay characteristics. However, there is currently great interest in near-space (17∼22 km) satellites called High Altitude Platforms (HAP) [6]. The delay characteristics for HAPs will be functional for terrestrial application. HAPs will also be characterized by wide coverage, offering reasonable coverage overlays for IMT-Advanced networks. Already, the ITU-R has issued the minimum performance requirements for HAPs providing 3G service in certain regions [7].

19.3 Cognitive Radio and Dynamic Spectrum

Software-Defined Radios (SDR) were initially defined so as to facilitate changing the characteristics and capabilities of a radio interface simply through reprogramming. Its evolution, Cognitive Radio (CR), was one where the programmability of the SDR can be made over-the-air and on-the-fly. What is more, however, is that a CR had sufficient processing capability to autonomously understand and react to various elements of the radio's context of operation [8]. Among other things,

these characteristics include identifying whether the current spectrum band of operation is the best spectrum available for the radio's active communication, and whether there are bands that are available and, for example, would offer greater bandwidth or better transmission quality. For CR to perform, it requires more than simply identifying whether or not a particular spectrum band is busy – rather, it becomes important that the radio recognizes what entity is utilizing that spectrum, and know for how long will this utilization will take place.

Such distinction is greatly important, especially in light of recent international cooperation between the different Telecommunications Regulatory Authorities (TRA) of the different countries and the ITU-R. These cooperations are at spectrum harmonization, refarming and reallocation. In addition, many countries now recognize primary and secondary users for certain bands, allowing for cooperative arrangements and coexistences between the different spectrum users, both licensed and unlicensed. It thus becomes possible for a secondary user to utilize spectrum "holes" or "empty spots" in a primary user's band or, depending on the band and mode of communication, for both primary and secondary users to operate in the same band [9]. Such cognition, however, is not limited to licensed bands. Bluetooth, for example, is already instilled with adaptability so as to overcome from other devices in the ISM band such as WiFi network elements or microwaves.

19.4 Network Intelligence

Services utilizing network and location analytics are already emerging in the smartphone applications market. Meanwhile, the proliferation of various sensing and actuating platforms, for example, ANT+ and IQRF, that interface directly with mainstream smartphone and network access types will soon allow for more valuable services that are more prompt, reliable and relevant. In this interweaved connectivity between context and personal preferences (both through settings and through non-invasive profiling), in addition to the service infrastructure of social networking platforms, the users' wireless and mobile experience will become much more enhanced. Another dimension of interest is that of utilizing network information to discern physical properties. Many examples of this have been displayed, both in research and industry. One of the commercial examples involves utilizing network traffic levels in recognizing actual street congestions [10].

For the considerations of access network operation, however, functionalities that employ network analytics include instilling reliable wireless communication, interference management and mitigation, power management, resource allocation, and reduced energy. Both LTE and WiMAX support various mechanisms for autonomous operation of network entities, and have made provisions for self-optimization in various aspects of their respective standards. For example, the operation of femtocells cannot do without autonomy, especially given the ad hoc nature of their deployment. Another example involves the required processing capabilities for Coordinated Multipoint Transmission (CoMP), which is one of the enabling technologies discussed in Chapter 2. As will be discussed next,

recognition of device usage patterns can also lead to great savings in network energy requirements.

It should be noted that an important aspect of instilling autonomous operation in network operation is motivated by several factors, chief among which is the physical interruption of operator personnel and administrators. Such self-management functionalities will also result in substantial reductions in signal and bandwidth requirements – a major cause of bandwidth and processing losses in traditional cellular networks [11].

19.5 Access Network Architecture

The introduction of 3GPP's X2 interface marked a particular evolutionary step in the design of access network infrastructure. Traditionally, base stations were connected to network cores in centralized star configuration, with each base station directly and independently connected to the access core. Such configuration, exercised up until the earlier releases of UTRAN, results in substantial handover latencies, especially when it came to IP-based mobility. Similarly with WiMAX, which is neutral to the choice of network core, support has been made to realizing flat architectures.

A direct advantage of flat architecture is greatly reduced handover latency times, which was mandated by the IMT-Advanced requirements letter. This advantage, consequently, results in reduced disruptions for multimedia IP handover as the users traverse the network [12]. Through internetworking base stations, user context can be transported from a serving base station to the target one without having to go back to the network core. As was observed, additional optimizations are also possible in instances where the user terminal moved between a base station and its children relay stations.

Careful network design, however, is required in order to achieve these desirable characteristics. Design considerations would include aspects such as where is it best to connect the access network to the core or the identifying topology configurations that match the projected traffic load while achieving certain levels of reliability. Looking beyond IMT-Advanced networks, interest has already started in what is called "ultra-flat architectures", wherein substantial processing is migrated from the network core to the network edges – the base stations [13]. Such migration, however, will largely depend on substantial advances taking place not in terms (of) processing capabilities, but also in inference frameworks. In such instances, the issues such as identifying the best location for a certain functionality, become more prominent.

19.6 Radio Resource Management

Radio resource management (RRM) functionalities oversee the allocation and maintenance of network resource to the various devices during network operation. RRM functionalities in IMT-Advanced comprise both traditional

and emerging modules, including modules for admission control, scheduling, resource reservation (for various prioritization objectives), spectrum management, ARQ/HARQ, and routing. The various modules comprise different elements of an overall framework, and are expected to operate in a cohesive manner, serving specific overall operational objectives. Designing frameworks for IMT-Advanced networks, however, is not without challenges. By requirements (no s?), IMT-Advanced networks are expected to deal with certain characteristics (,) among which are an immense magnitude of traffic from both users and devices, a range of traffic requirements for various services and applications, a range of mobility speeds, and different types of access technologies and modes. A definite problem of traditional framework designs is that they do not scale.

Complexity, hence, becomes a key issue to overcome when designing such frameworks, and one that is prominent at the different levels of network management. For example, the difficulty of scheduling multi-carrier access techniques, both OFDMA and SC-FDMA, was illustrated in Chapter 17. And while the separation of the time and frequency aspects of resource allocations does lead to significant operational optimization, scheduling becomes more cumbersome when introducing advances such as MIMO, either at the single cell or the multiple cell level [14]. Another example of complexity can be found at a higher management level, and has to do with admission control of connections or flows. IMT-Advanced networks will employ different modes of operation, including point-to-multipoint, where a base station communicates directly to the device, relaying where the base stations communicate with the devices through one or more relay stations, or femtocells where the devices connect through the Internet. Meanwhile, IMT-Advanced networks will support access heterogeneity, which adds the selection of access technology to the possible connection choices. In addition, the flexibility in spectrum allocations will also make possible varying the spectrum band through which the device is connected, that is, a spectrum handover. Considering that more than 50 Billion devices will be connected in the future, the importance of simplifying network selection mechanisms becomes more pressing [15].

This complexity issue has already been tackled in several ways. For example, the above noted notion of small cells “opens up” the capacities at the network end – a strong leverage when considering different connection possibilities. At the same time, the introduction of flat architectures have also simplified the considerations of the RRM as they have forced the decision making to be more localized, focusing only at users within the cell and the technologies overlaying the cell’s coverage. Within the research, much work has addressed the possibility of Common RRM, whereby the resources of overlaid access technologies can be jointly managed – a powerful advance that is viable for technologies administered by a single operator. Advances are expected in the AAA that would further facilitate inter-operator resource agreements and management. These advances, however, will take a longer time to realize.

Nevertheless, there are certain fundamental aspects of RRM design that need to be highlighted [16]. One is that a tradeoff exists between value – not performance – optimization and the amount of information, and consequently the signaling, required to achieve that optimization. For example, up-to-date information about the location and application requirements of different users connected through different access technologies can be made to be promptly available at a central entity. A variant case of this setting would be the one encountered in CoMP transmissions. The tradeoff entails that while better allocations can be made with prompt user and medium information, an acceptable performance can be achieved with some of this information delayed or missing. This raises another important issue, and that is where is it best to locate this decision making entity. The problem of finding this location should not be decoupled from the one encountered in designing the access network's flat architecture. Another fundamental aspects is concerned with how IMT-Advanced networks will ultimately be delivering Internet traffic and services. End-to-end performance therefore plays a substantial role that is equal to the access level performance. And while advances such as deep packet inspection will soon materialize IETF-based QoS (DiffServ, IntServ, MPLS) in cellular access networks [17], inter-domain optimization remains an outstanding challenge.

19.7 Green Wireless Access

By some estimates, cellular networks consume 0.5 % of world-wide energy consumption, with 1 % consumed by the user handsets and 99 % consumed by the network [18]. Meanwhile, multiple-interface phones (Cellular with WiFI, Bluetooth, ANT+, etc.) have been observed to deplete their batteries much faster when all the radios are active all the time. Not surprisingly, then, that several initiatives and research projects have focused on reducing the energy requirements of wireless and mobile networks over the past few years. The projects, in general, vary in their approaches and their objectives. Some, for example, have focused on energy reduction through interference management – reducing the energy requirements of mobile handsets to reliably transmit its data. Network design plays an important role, whereby the location of the fixed base stations and the trajectory of the mobile stations are decided in a manner that also reduces handset energy expenditure. Meanwhile, energy can definitely be added to the considerations of network selection. Advances in dynamic spectrum allocation will also play a major role.

These enhancements, however, focus on handset energy expenditure. To alleviate some of the network expenditure, it is possible (to) utilize renewable energy sources such as solar and wind turbines. More advanced mechanisms, however, can also be employed. For example, it is possible to deploy high density access configurations whereby the all base stations would be turned in instances of high

demand, and only a portion of the base stations would operate when the demand decreases. Naturally, a small coverage would be used when all base stations are turned on, and a wider coverage when only a portion is operating.

As in the case with the design of RRM frameworks, there are certain tradeoffs bound to how "green" the operation of a wireless network can be [19]. These include the tradeoff between deployment efficiency and energy efficiency, where deployment efficiency refers to the network throughput per cost performance vs. the network's energy consumption. There is also the tradeoff between spectrum efficiency and energy efficiency – directly relevant to the optimization-overhead tradeoff discussed above. Spectrum efficiency, particularly, is an energy-exhaustive process, as it requires sensing in several spectrum bands, possibly simultaneously. Such sensing also needs to be made during secondary user transmission, as secondary users are required to vacate the primary user's spectrum once the latter begins communicating. The remaining tradeoffs include the bandwidth vs. power and delay vs. power tradeoffs. These tradeoffs, while open for optimizations, should be minded in the design of green networks.

References

[1] See "Evaluation Reports" at http://www.itu.int/en/ITU-T/gsi/iot/Pages/default.aspx.
[2] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010-2015", Whitepaper, February 2011, available at http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html.
[3] UN, Department of Economic and Social Affairs – Population Division, "World Population to 2300", 2004. (available at http://www.un.org/esa/population/publications/longrange2/WorldPop2300final.pdf).
[4] L.M. Ericsson, "More than 50 Billion Connected Devices," February 2011.
[5] M. Dohler et al., "Is the PHY Layer Dead?," *IEEE Communications Magazine*, Volume 49, Issue 4, pp. 159–65, April 2011.
[6] S. Karapantazi and F. Pavlidou, "Broadband Communications visa High-Altitude Platforms: A Survey," *IEEE Surveys and Tutorials*, Volume 7, Issue 1, pp. 2–31, First Qtr. 2005.
[7] Recommendation ITU-R M.1456, "Minimum performance characteristics and operational conditions for high altitude platform stations providing IMT-2000 in the bands 1 885-1 980 MHz, 2 010-2 025 MHz and 2 110-2 170 MHz in Regions 1 and 3 and 1 885-1 980 MHz and 2 110-2 160 MHz in Region 2", http://www.itu.int/rec/R-REC-M.1456-0-200005-I/en.
[8] J. Mitola and G.Q. Mguire, Jr., "Cognitive Radio: Making Software Radios More Personal," *IEEE Personal Communications*, Volume 6, Issue 4, pp. 13–18, August 1999.
[9] S. Haykin, "Cognitive Radio: Brain-Empowered Wireless Communications," *IEEE Journal on Selected Areas in Communications*, Volume 23, Issue 2, pp. 201–20, February 2005.
[10] The Metro Traffic Engine by the Intelligent Mechatronic Systems, http://www.intellimec.com/traffic/.
[11] FierceWireless's panel, "The Pros and Cons and of Diverting Mobile Data Traffic," available at http://www.fiercewireless.com/webinars/pros-and-cons-diverting-mobile-data-traffic.
[12] D. Amzallag, J.S. Naor and D. Raz, "Algorithmic Aspects of Access Networks Design in B3G/4G Cellular Networks", in *Proceedings of the 26th IEEE International Conference on Computer Communications*, pp. 991–9, May 2007.
[13] L. Bokor, Z. Faigl and S. Imre, "Flat Architectures: Towards Scalable Future Internet Mobility," *Lecture Notes on Computer Science*, J. Domingue et al. (eds): *Future Internet Assembly*, Volume 6656/2011, pp. 35–50, 2011.

[14] A. Maeder and N. Zein, "OFDMA in the Field: Current and Future Challenges," *ACM SIGCOMM Computer Communications Review*, Volume 40, Issue 5, pp. 71–6, October 2010.
[15] D.E. Charilas and A.D. Panagopoulos, "Network Selection Problem: Multiaccess Radio Network Environments," *IEEE Vehicular Technology Magazine*, Volume 5, Issue 4, pp. 40–9, December 2010.
[16] J. He, J. Rexford and M. Chiang, "Don't Optimize Existing Protocols, Design Optimizable Protocols," *ACM SIGCOMM Computer Communication Review*, Volume 37, Issue 3, pp. 53–8, July 2007.
[17] L. Jorguseki, "Vision on Radio Resource Management (RRM) and Quality of Service (QoS) for Wireless Communication Systems in Year 2020", *Globalization of Mobile and Wireless Communications*, R. Prasad et al. (eds) Springer, Netherlands, 2011.
[18] Going Greener, Vodafone, http://www.vodafone.com/content/index/uk_corporate_responsibility/greener.html.
[19] Y. Chen et al., "Fundamental Tradeoffs on Green Wireless Networks," *IEEE Communications Magazine*, 2011.

Index

LTE, LTE-Advanced and WiMAX: Towards IMT-Advanced Networks, First Edition.
Abd-Elhamid M. Taha, Najah Abu Ali and Hossam S. Hassanein.