LTE-Advanced

A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies

Sassan Ahmadi







Samsung Ex. 1005. Page 1 of 1147

LTE-Advanced

This page intentionally left blank

LTE-Advanced

A Practical Systems Approach to Understanding the 3GPP LTE Releases 10 and 11 Radio Access Technologies

Sassan Ahmadi



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier
The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

Copyright © 2014 Elsevier Inc. All rights reserved

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher.
Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+ 44) (0) 1865 843830; fax (+ 44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively, visit the Science and Technology Books website at www.elsevierdirect.com/rights for further information.

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-405162-1

For information on all Academic Press publications visit our website at elsevierdirect.com

Typeset by MPS Limited, Chennai, India
www.adi-mps.com

Printed and bound in United States of America

14 15 16 10 9 8 7 6 5 4 3 2 1



Contents

Preface xv
Acknowledgments xix
Abbreviations and Acronyms xxi

CHAPTER 1 Introduction to LTE-Advanced 1
1.1 Background on IMT-Advanced standards development 5
1.2 Requirements for the IMT-Advanced systems 8
1.3 Introduction to 3GPP standards 11
1.4 Evolution of 3GPP standards 18
1.5 A perspective on the future trends in cellular networks 20
References 26
3GPP Specifications 27
Articles 27

CHAPTER 2 Network Architecture 29
2.1 Overall network architecture 31
2.1.1 General concepts 31
2.1.2 Evolution of the network architecture 38
2.2 EPC architecture 42
2.2.1 Reference model and network interfaces 42
2.2.2 Public land mobile network 43
2.2.3 Packet data network gateway 46
2.2.4 Serving gateway 47
2.2.5 Mobility management entity 47
2.2.6 Policy and charging rules function 48
2.2.7 Home subscriber server 49
2.2.8 Authorization, authentication, and accounting 49
2.2.9 EPC identifiers 50
2.3 Roaming architecture and interworking with other networks 50
2.4 E-UTRAN architecture 57
2.4.1 E-UTRAN functions 59
2.4.2 E-UTRAN interfaces 59
2.4.3 E-UTRAN identifiers 60
2.4.4 Support of HeNB in E-UTRAN 61
2.4.5 Support of RRHs in E-UTRAN 63
2.4.6 Support of relays in E-UTRAN 64
2.5 E-MBMS architecture 67
2.6 Positioning network architecture 71
2.7 Mobility management 74

Samsung Ex. 1005, Page 6 of 1147

2.8 Radio resource management 75
2.9 Quality of service/quality of experience 77
2.10 Security in E-UTRAN 78
2.11 Support of voice and messaging services in LTE/LTE-Advanced 80
2.11.1 IMS architecture 80
2.11.2 Single radio voice call continuity 84
2.11.3 Voice over LTE 90
2.11.4 Circuit-switched fallback 94
2.12 LIPA, selected IP traffic offload, and IP flow mobility 100
2.12.1 Local IP access 101
2.12.2 Selected IP traffic offload 105
2.12.3 IP flow mobility 107
2.12.4 Wi-Fi offloading 110
2.13 Mobile IP 111
2.13.1 Client-based IP mobility 113
2.13.2 Network-based IP mobility 113
2.14 Network architecture evolution and small cells 114
References 116

CHAPTER 3 E-UTRAN and EPC Protocol Structure 121
3.1 EPS reference model and protocol structure 124
3.1.1 Functional split between E-UTRAN and EPC 127
3.1.2 Protocol stack for S1 user-plane 129
3.1.3 Protocol stack for S1 control-plane 129
3.1.4 Protocol stack for S5/S8 interface 131
3.1.5 Protocol stack for S11 interface 131
3.1.6 Protocol stack for X2 user-plane 132
3.1.7 Protocol stack for X2 control-plane 134
3.2 E-UTRAN and NAS protocols 135
3.3 LTE-Uu interface protocols 138
3.4 Support for relay nodes 144
3.5 Support for HeNB nodes 148
References 151

CHAPTER 4 System Operation and UE States 153
4.1 General UE behavior 154
4.2 UE initialization procedure (cell search and cell selection) 157
4.3 EPS connection and mobility management 166
4.3.1 EPS mobility management 167
4.3.2 EPS connection management 168
4.4 UE states 169
4.4.1 Idle mode procedures 171
4.4.1.1 Tracking area update 174

4.4.1.2 Paging 174
4.4.1.3 Cell reselection 176
4.4.1.4 Idle mode DRX 177
4.4.2 Connected mode procedures 178
4.5 Network attach procedures 180
4.5.1 Random access procedure 185
4.5.2 RRC connection establishment, modification, and release 188
4.5.3 Service establishment and QoS 191
4.6 Data call establishment procedures 194
4.6.1 UE-originated data call procedure 194
4.6.2 UE-terminated data call procedure 196
4.7 Mobility procedures 198
4.7.1 Idle mode mobility procedures 201
4.7.2 Connected mode mobility procedures 202
4.7.3 HetNet mobility 203
4.7.4 S1-based handover procedure 208
4.7.5 X2-based handover without S-GW change 212
4.7.6 X2-based handover with S-GW change 214
4.7.7 Inter-RAT mobility procedures 216
4.8 Overview of IP packet processing procedures 222
References 224

CHAPTER 5 Radio Resource Control Functions 227
5.1 RRC states and state transitions 230
5.1.1 RRC_CONNECTED state 232
5.1.2 RRC_IDLE state 234
5.2 RRC connection management 234
5.2.1 Signaling and data radio bearers 240
5.2.2 Quality of service control 244
5.3 RRC mobility management 246
5.3.1 Mobility in RRC_CONNECTED state 251
5.3.2 Mobility in RRC_IDLE state 252
5.3.3 Intra-RAT mobility and handover 253
5.3.4 Inter-RAT mobility and handover 255
5.4 Measurements and measurement reporting 257
5.5 RRC procedures 262
5.5.1 System information 262
5.5.2 Paging 266
5.5.3 PLMN and cell selection 275
5.5.4 Cell reselection 278
5.5.5 Radio link failure 281
5.6 RRC timers and constants 282
5.7 RRC message structure 282
References 286

CHAPTER 6 Packet Data Convergence Protocol Functions 289
6.1 Ciphering and integrity protection functions 294
6.2 Header compression 296
6.3 Robust header compression 299
6.4 PDCP PDU packet formats 303
6.5 LTE security aspects 304
References 308

CHAPTER 7 Radio Link Control Functions 311
7.1 RLC architecture 314
7.2 RLC transfer modes 316
7.2.1 Transparent mode 317
7.2.2 Unacknowledged mode 319
7.2.3 Acknowledged mode 321
7.3 RLC ARQ procedure 327
7.3.1 ARQ principles 327
7.3.2 ARQ operation 330
7.4 RLC PDU packet formats 332
7.4.1 Transparent mode PDU format 332
7.4.2 Unacknowledged mode PDU formats 334
7.4.3 Acknowledged mode PDU formats 335
7.4.4 Status PDU format 336
7.5 RLC variables, constants, and timers 337
References 340

CHAPTER 8 Medium Access Control Functions 343
8.1 MAC architecture and services 345
8.2 Logical and transport channel mapping 349
8.3 Downlink/uplink data transfer 353
8.3.1 HARQ principles 354
8.3.2 Downlink HARQ timing and signaling protocols 357
8.3.3 Uplink HARQ timing and signaling protocols 358
8.4 DRX operation 366
8.5 MAC control elements 371
8.5.1 Uplink timing advance control element 372
8.5.2 BSR control element 374
8.5.3 Power headroom control element 375
8.5.4 Extended power headroom control element 376
8.5.5 C-RNTI control element 377
8.5.6 Activation/deactivation control element 378
8.5.7 MCH scheduling information control element 378
8.5.8 UE contention resolution identity control element 379

8.6 MAC PDU packet formats 380
8.7 MAC scheduling services 381
8.7.1 Dynamic and semi-persistent scheduling 384
8.7.2 Proportional fair scheduling algorithm 386
8.8 RNTI mapping 387
8.9 Random access procedure 389
8.9.1 Contention-based RACH procedure 392
8.9.2 Contention-free RACH procedure 392
References 396

CHAPTER 9 Downlink Physical Layer Functions 399
9.1 Overview of downlink physical layer processing 402
9.2 Characteristics of wireless channels 404
9.3 LTE/LTE-Advanced operating frequencies and band classes .411
9.4 Principles of OFDM 419
9.4.1 Peak-to-average power ratio of OFDM signals 426
9.4.2 Effect of non-linearity on OFDM signals 428
9.4.3 Effect of carrier frequency offset on OFDM signals 430
9.4.4 Effect of oscillator phase noise on OFDM signals 434
9.5 Multiple access schemes 437
9.6 Duplex modes 440
9.7 Frame structure 445
9.7.1 Frame structure type 1 446
9.7.2 Frame structure type 2 447
9.7.3 MBSFN subframe structure 452
9.8 Resource structure, allocation, mapping 453
9.8.1 Physical and virtual resource blocks 454
9.8.2 Resource allocation 460
9.8.2.1 Resource allocation type 0 462
9.8.2.2 Resource allocation type 1 463
9.8.2.3 Resource allocation type 2 465
9.8.2.4 PRB bundling 467
9.8.2.5 Resource element group 467
9.9 Time and frequency synchronization 471
9.9.1 Primary synchronization signals 473
9.9.2 Secondary synchronization signals 474
9.10 Downlink reference signals and channel estimation 478
9.10.1 Reference signal design criteria 480
9.10.2 Pilot-based channel estimation 486
9.10.3 Cell-specific reference signals 491
9.10.3.1 Received signal strength indicator measurement 496

9.10.3.2 Reference signal received power measurement 496
9.10.3.3 Reference signal received quality measurement 496
9.10.4 UE-Specific reference signals associated with PDSCH 497
9.10.5 CSI reference signals 502
9.10.6 Enhanced PDCCH demodulation reference signals 512
9.10.7 Positioning reference signals 514
9.10.8 MBSFN reference signals 516
9.10.9 Reference signal overhead calculation 518
9.11 Channel coding and modulation 519
9.11.1 Principles of convolutional coding 521
9.11.2 Principles of turbo coding 527
9.11.3 Baseband digital modulation schemes 533
9.11.4 Channel coding, multiplexing, and interleaving 537
9.12 Downlink physical channel processing 546
9.12.1 Physical control format indicator channel 553
9.12.2 Physical broadcast channel 556
9.12.3 Physical hybrid-ARQ indicator channel 560
9.12.4 Physical downlink control channel 568
9.12.4.1 PDCCH physical layer processing 568
9.12.4.2 DCI formats 574
9.12.4.3 Common and UE-specific search spaces and blind decoding 576
9.12.5 Enhanced PDCCH 597
9.12.6 Physical downlink shared channel 604
9.13 Principles of multi-antenna transmission 608
9.13.1 Capacity of MIMO channels 610
9.13.2 Instantaneous capacity, ergodic capacity, and outage capacity of MIMO systems 617
9.13.3 Spatial multiplexing and diversity 620
9.13.4 MIMO receivers 624
9.13.4.1 Maximum likelihood receiver 624
9.13.4.2 Successive interference cancelation receiver 624
9.13.4.3 Zero-forcing receiver 625
9.13.4.4 Minimum mean square error receiver 626
9.13.4.5 Maximal ratio combining receiver 627
9.13.4.6 MMSE-interference rejection combining receiver 628
9.13.4.7 Sphere decoding 630
9.13.4.8 CSI feedback 632

9.13.5 Precoding and beamforming 636
9.13.6 SU-MIMO and MU-MIMO 642
9.13.7 Collaborative MIMO and collaborative spatial multiplexing 652
9.13.8 Non-linear precoding and dirty-paper coding 654
9.14 LTE/LTE-Advanced downlink multi-antenna transmission schemes 661
9.14.1 Transmit diversity 666
9.14.2 Spatial multiplexing 668
9.14.3 Cyclic delay diversity 672
9.14.4 Beamforming and MU-MIMO 674
9.15 PDSCH transmission modes 679
9.16 OFDM signal generation 686
9.17 Timing advance 686
9.18 Power control and link adaptation 690
9.18.1 Basic concepts 691
9.18.2 Downlink power allocation 698
9.18.3 PUSCH power control 700
9.18.4 PUCCH power control 703
9.18.5 SRS power control 704
9.18.6 PRACH power control 705
9.18.7 Power headroom 706
9.18.8 Link adaptation 707
References 712

CHAPTER 10 Uplink Physical Layer Functions 721
10.1 Overview of uplink physical layer processing 724
10.2 Principles of SC-FDMA 724
10.3 SC-FDMA and clustered SC-FDMA 734
10.4 Uplink frame structure 736
10.5 Physical resource structure and channel mapping 738
10.5.1 Physical resource blocks 742
10.5.2 Uplink resource allocation 743
10.5.2.1 Uplink resource allocation type 0 747
10.5.2.2 Uplink resource allocation type 1 748
10.6 Uplink reference signals and channel estimation 753
10.6.1 Uplink demodulation reference signals 754
10.6.2 Sounding reference signals 762
10.7 Uplink physical channel processing 769
10.7.1 Physical uplink control channel 774
10.7.2 PUCCH formats 778
10.7.2.1 PUCCH formats 1, 1a, 1b 780
10.7.2.2 PUCCH formats 2, 2a, 2b 785
10.7.2.3 PUCCH format 3 797

10.7.3 Physical uplink shared channel 802
10.7.4 Physical random-access channel 821
10.8 SC-FDMA signal generation 837
10.9 LTE/LTE-Advanced uplink multi-antenna transmission schemes 838
10.9.1 Transmit antenna selection 839
10.9.2 Virtual MU-MIMO (cooperative MIMO) 840
10.9.3 Closed-loop spatial multiplexing 841
10.9.4 Transmit diversity 844
10.9.5 Uplink MIMO receivers 846
10.9.5.1 LMMSE receiver 849
10.9.5.2 MRC receiver 849
10.9.5.3 MMSE-IRC receiver 849
10.9.5.4 SIC receiver 850
10.10 Uplink transmission modes 851
10.11 Uplink feedback in LTE/LTE-Advanced 852
10.12 Uplink transmitter and receiver chain 861
References 866

CHAPTER 11 Link-Level and System-Level Performance of LTE-Advanced 875
11.1 Definition of the performance metrics 882
11.1.1 Link budget 882
11.1.2 User and sector throughput 883
11.1.3 Outage throughput 885
11.1.4 Packet delay 885
11.1.5 Spectral efficiency 886
11.1.6 Jain fairness index 886
11.1.7 Geometry (wideband SINR distribution) 887
11.1.8 Effective interference over thermal 887
11.1.9 Cell spectral efficiency and cell-edge user spectral efficiency 888
11.2 Calculation of static and dynamic overhead 889
11.3 Traffic models 891
11.3.1 Statistical model for conversational speech 891
11.3.2 Full-buffer traffic model 893
11.4 Link-to-system mapping (PHY abstraction) 893
11.5 IMT-Advanced test environments 902
11.6 Network layout for system-level simulations 902
11.7 IMT-Advanced evaluation methodology and baseline configurations 904
11.8 Link-level and system-level channel models 904
11.9 LTE-Advanced link-level and system-level performance 916
11.9.1 Cell spectral efficiency and cell-edge spectral efficiency 916

11.9.2 VoIP capacity 922
11.9.3 Mobility 923
11.9.4 Peak spectral efficiency 925
11.9.5 User-plane/control-plane latency and handover interruption time 925
11.9.6 Estimation of the L1/L2 overhead 932
11.9.7 Link budget 936
11.9.8 Evaluation of LTE-Advanced against LTE Rel-10 requirements 936
11.10 Miscellaneous evaluations 936
References 947

CHAPTER 12 Coordinated Multipoint Transmission and Reception (CoMP) 951
12.1 Interference cancellation in cellular networks 952
12.2 Distributed antenna systems 955
12.3 Theory of CoMP operation 959
12.4 CoMP architectures 964
12.4.1 Air-interface and network support for CoMP schemes 968
12.4.2 Downlink CoMP 969
12.4.3 Uplink CoMP 971
12.4.4 CSI measurement and feedback in CoMP-enabled networks 974
12.5 Effects of X2 latency on CoMP performance 977
12.6 Performance of CoMP scenarios 978
References 981

CHAPTER 13 Carrier Aggregation 985
13.1 Principles of carrier aggregation 986
13.2 Deployment scenarios 990
13.3 Physical and MAC layer aspects of carrier aggregation 992
13.3.1 Physical layer aspects 992
13.3.2 MAC layer aspects 999
13.4 Protocol and signaling aspects of carrier aggregation 1002
13.5 Radio resource management aspects of carrier aggregation 1005
13.6 RF and implementation aspects of carrier aggregation 1006
13.6.1 Temporal/spectral masks and unwanted emissions 1008
13.6.2 Operating bands for carrier aggregation 1010
13.6.3 Transmitter characteristics for carrier aggregation 1016
13.6.4 Receiver characteristics for carrier aggregation 1022
References 1027

CHAPTER 14 Enhanced Inter-cell Interference Coordination and Multi-radio Coexistence 1029
14.1 Intra-cell/inter-cell interference mitigation methods 1030
14.2 Heterogeneous networks 1035
14.3 Enhanced inter-cell interference coordination (eICIC) 1040
14.3.1 CA-based eICIC schemes 1041
14.3.2 Non-CA-based eICIC 1044
14.3.3 Further enhanced ICIC 1048
14.4 Multi-radio coexistence and interference avoidance 1051
14.4.1 Collocated multi-radio coexistence issues 1052
14.4.2 In-device coexistence scenarios 1054
14.4.3 In-device coexistence and interference avoidance 1056
References 1067

CHAPTER 15 Positioning and Multimedia Broadcast/Multicast Services 1069
15.1 Positioning and location-based services 1071
15.1.1 E-UTRA positioning network architecture 1074
15.1.2 Network-based and UE-based positioning methods 1078
15.1.3 Positioning methods in LTE/LTE-Advanced 1080
15.1.3.1 Principles of GPS 1081
15.1.3.2 A-GNSS positioning 1085
15.1.3.3 Observed time difference of arrival positioning 1086
15.1.3.4 E-CID-based positioning 1088
15.1.3.5 UTDOA positioning 1089
15.2 Evolved multimedia broadcast/multicast service 1090
15.2.1 E-MBMS network architecture 1091
15.2.2 Basic definitions 1094
15.2.3 E-MBMS user-plane and control-plane protocols 1094
15.2.4 Physical and MAC layer aspects of E-MBMS 1095
15.2.5 E-MBMS single frequency network transmission 1098
15.2.6 Mobility support in E-MBMS 1102
15.2.7 E-MBMS counting 1103
References 1104

Index 1107

Preface

The past two decades have witnessed a phenomenal paradigm shift in the design methodology, and substantial improvement in the performance of mobile broadband wireless access technologies. With the increasing consumer demand for diverse wireless multimedia and data applications, and mobile Internet, improved quality and increasing capacity of the wireless networks have become a priority for the standards development organizations such as 3GPP. 3GPP started the development of the evolved packet system in late 2004, which subsequently resulted in specification and standardization of a new OFDM-based radio access network and an IP-based core network in 3GPP Rel-8, which was further enhanced in the subsequent releases. LTE Rel-10 (i.e., LTE-Advanced) was submitted as a candidate to ITU-R by 3GPP, and was selected as an IMT-Advanced technology in October 2010. The studies and research have continued in order to further advance and improve radio access technologies beyond LTE Rel-10 and HSPA Rel-10 by development of 3GPP Rel-11. The next generation of radio access technologies, informally referred to as systems beyond IMT-Advanced, are currently under investigation. Some of the main considerations in the design and development of new radio interfaces include higher spectral efficiency by using advanced multi-antenna technologies, carrier aggregation, distributed antenna systems, heterogeneous networks and small cells, energy efficiency and green radios, as well as the use of cognitive radios and software-defined radios for more efficient and flexible use of the radio spectrum. The systems beyond IMT-Advanced will encompass the capabilities of previous generations, as well as new communication schemes such as machine-to-machine, machine-to-person, and person-to-machine.

The framework for development of IMT-Advanced and systems beyond IMT-Advanced can be viewed from multiple perspectives including the users, manufacturers, application developers, network operators, and service and content providers. From the user's perspective, there is a demand for a variety of services, content, and applications whose capabilities will increase over time. The users expect services to be ubiquitously available through a variety of delivery mechanisms and service providers using a variety of wireless devices. From a service provision perspective, the domains share some common characteristics. Wireless service provision is characterized by global mobile access (terminal and personal mobility), improved security and reliability, higher service quality, and access to personalized multimedia services, Internet, and location-based services via one or multiple user terminals. Multi-radio operation requires seamless interaction

Samsung Ex. 1005, Page 16 of 1147

among systems in order for the user to be able to receive/transmit a variety of content via different delivery mechanisms, depending on the device capabilities, location, and mobility, as well as user profile. Different radio access systems can be interconnected via a flexible core network and appropriate interworking functions. In this way, a user can be connected through different radio access systems to the network and utilize the services. The interworking between different radio access systems in terms of horizontal or vertical handover and seamless connectivity with service negotiation, mobility, security, and quality of service management are key requirements of radio-agnostic networks.

Similarity of services and applications across different radio access systems is beneficial not only to users, but also to network operators and content providers, stimulating the current trend toward convergence. Furthermore, similar user experience across different radio access systems leads to large-scale adoption of products and services, common applications, and content. Access to a service or an application may be performed using one system, or may be performed using multiple systems simultaneously. The increasing predominance of IP-based applications has been a key driver for the convergence trend in the core network and access network technologies.

The evolution of legacy IMT-2000 systems and the IMT-Advanced systems has employed several new concepts and functionalities, including adaptive modulation and coding and link adaptation, OFDM-based multiple access schemes, single-/multi-user multi-antenna concepts and techniques, dynamic QoS control, mobility management and handover between heterogeneous radio interfaces (vertical and horizontal), robust packet transmission, error detection and correction, multi-user detection, and interference cancellation. Systems beyond IMT-Advanced may further utilize sophisticated schemes including software-defined radio and reconfigurable RF and baseband processing, adaptive radio interface, mobile ad hoc networks, routing algorithms, and cooperative communication.

This book provides an in-depth description of 3GPP Rel-10/11 standards from a systems engineering point of view. A unique top-down systems approach is used to identify and describe the functional and protocol components of the 3GPP Rel-10/11 radio access technologies, starting from the radio access and core network architecture and then continuing with the theory of operation, interfaces, and practical implementation and deployment considerations. The book contains the background theoretical information about the operation of mobile stations, base stations, and other radio interface nodes. The distinctive features of this book relative to similar publications include an organized and systems engineering approach, self-containment of the technical topics discussed in the book, in-depth description of the theoretical concepts, real-life implementation issues, and practical considerations using clear and unambiguous block diagrams, state diagrams, and signal flow graphs.

Many articles, blogs, white papers, application notes, book chapters, and books have been published on the subject of LTE/LTE-Advanced and fourth generation of cellular networks since 2004, varying from academic theses to network

operator analyses and manufacturers' application notes. By their very nature, these publications have viewed these subjects from one particular perspective whether it is academic, operational, or promotional/commercial. A very different and unique approach has been taken in this book, and that is a top-down systems approach to understanding the system operation and the design principles of the underlying functional components of fourth generation radio access networks. This book can be considered as the most up-to-date and complete technical reference for the design of LTE/LTE-Advanced systems. In this book, the protocol layers and functional elements of LTE/LTE-Advanced radio access and core networks are described.

This book is a reflection of the author's several years of participation in the inception, design and development, standardization, and implementation of LTE/LTE-Advanced, as well as numerous books, articles, white papers, public presentations, public blogs and online notes, simulations, etc., that the author has come across and studied since 2004 about LTE/LTE-Advanced, along with the author's 25 + years of experience in communication theory, signal processing, systems engineering, and cellular communication.

Samsung Ex. 1005, Page 18 of 1147

This page intentionally left blank

Acknowledgments

The author would like to acknowledge and thank his friends in 3GPP, 3GPP2, ITU-R, and IEEE for their encouragement, consultation, and assistance in selection of the content, as well as editing and improving the quality of the chapters.

The author would like to sincerely appreciate the Academic Press (Elsevier) publishing and editorial staff for providing the author with the opportunity to publish this work and for their assistance, cooperation, patience, and understanding throughout this project.

Finally, the author is immeasurably grateful to his family for their unwavering encouragement, cooperation, support, patience, and understanding throughout this project.

Samsung Ex. 1005, Page 20 of 1147

This page intentionally left blank

Abbreviations and Acronyms

3GPP	3rd generation partnership project
AAA	authorization, authentication, and accounting
ABS	almost blank subframe
AC	access class
ACIR	adjacent channel interference ratio
ACK	acknowledgment
ACLR	adjacent channel leakage ratio
ADC	analog to digital converter
AF	application function
AM	acknowledged mode
AMBR	aggregate maximum bit rate
AMC	adaptive modulation and coding
ANR	automatic neighbor relation
AoA	angle of arrival
AoD	angle of departure
AP	application protocol
APN	access point name
ARFCN	absolute ratio frequency channel number
ARP	allocation and retention priority
ARQ	automatic repeat request
AS	access stratum/angular spread
AWGN	additive white Gaussian noise
BC	broadcast channel (MU-MIMO)
BCCH	broadcast control channel
BCH	broadcast channel
BER	bit error rate
BPSK	binary phase shift keying
BSR	buffer status report
C/I	carrier-to-interference ratio
CA	carrier aggregation
CAZAC	constant amplitude zero autocorrelation
CBC	cell broadcast center
CBE	cell broadcast entity
CC	component carrier
CCE	control channel element

Samsung Ex. 1005, Page 22 of 1147

CCO	cell change order
CDD	cyclic delay diversity
CDF	cumulative distribution function
CDM	code division multiplexing
CFI	control format indicator
CFO	carrier frequency offset
CGI	cell global identifier
CIF	carrier indicator field
CM	cubic metric
CMAS	commercial mobile alert service
CMC	connection mobility control
CoMP	coordinated multi-point
CP	cyclic prefix
CQI	channel quality indicator
CRC	cyclic redundancy check
CRE	cell range extension
C-RNTI	cell RNTI
CRS	cell-specific reference signal
CSA	common subframe allocation
CSFB	circuit-switched fallback
CSG	closed subscriber group
CSG ID	closed subscriber group identity
CSI	channel state information
CSI-IM	CSI interference measurement
CSI-RS	CSI reference signal
CSS	CSG subscriber server
DAC	digital to analog converter
DAI	downlink assignment index
DCCH	dedicated control channel
DCI	downlink control information
DeNB	donor eNB
DFT-S-OFDM	DFT spread OFDM
DL	downlink
DL TFT	downlink traffic flow template
DL-SCH	downlink shared channel
DM-RS	demodulation reference signal
DoA	direction of arrival
DPC	dirty paper coding
DRB	data radio bearer
DRX	discontinuous reception
DS-CDMA	direct-sequence code division multiple access
DTCH	dedicated traffic channel
DTX	discontinuous transmission
DwPTS	downlink pilot time slot

EAB extended access barring
ECCE enhanced control channel element
ECGI E-UTRAN cell global identifier
E-CID enhanced cell-ID (positioning method)
ECM EPS connection management
ECN explicit congestion notification
eICIC enhanced inter-cell interference coordination
EMM EPS mobility management
eNB E-UTRAN node B/evolved node B
EPC evolved packet core
ePDCCH enhanced physical downlink control channel
ePHR extended power headroom report
EPRE energy per resource element
EPS evolved packet system
E-RAB E-UTRAN radio access bearer
eREG enhanced resource-element group
E-SMLC evolved serving mobile location center
ETWS earthquake and tsunami warning system
E-UTRA evolved UTRA
E-UTRAN evolved UTRAN
EVM error vector magnitude
FDD frequency division duplex
FDM frequency division multiplexing
FFT fast Fourier transform
FSTD frequency switched transmit diversity
FTP file transfer protocol
GBR guaranteed bit rate
GERAN GSM edge radio access network
GGSN gateway GPRS support node
GNSS global navigation satellite system
GP guard period
GPS global positioning system
GSM global system for mobile communication
GT guard time
GTP GPRS tunneling protocol
GUMMEI globally unique MME identifier
GUTI globally unique temporary identifier
GW gateway
HARQ hybrid ARQ
HeNB home eNB
HeNB GW home eNB gateway
HetNet heterogeneous network
HFN hyper frame number
HI HARQ indicator

HO	handover
HRPD	high rate packet data
HSPA	high speed packet access
HSS	home subscriber server
HTTP	hypertext transfer protocol
ICIC	inter-cell interference coordination
ID	identity
IDC	in-device coexistence
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IFFT	inverse fast Fourier transform
IMD	intermodulation distortion
IMS	IP multimedia subsystem
IoT	interference over thermal
IP	Internet protocol
IQ	in-phase/quadrature
IRC	interference rejection combining
ISI	inter-symbol interference
ISM	industrial, scientific, and medical
LB	load balancing
LBI	linked EPS bearer identifier
LCG	logical channel group
LCID	logical channel identifier
LCR	low chip rate
LCS	location service
L-GW	local gateway
LIPA	local IP access
LLR	log likelihood ratio
LMU	location measurement unit
LNA	low-noise amplifier
LPPa	LTE positioning protocol annex
LTE	long-term evolution
MAC	medium access control/multiple access channel (MU-MIMO)
MAC-I	message authentication code for integrity
MAP	maximum a posteriori probability
MBMS	multimedia broadcast/multicast service
MBR	maximum bit rate
MBSFN	multimedia broadcast multicast service single frequency network
MCCH	multicast control channel
MCE	multi-cell/multicast coordination entity
MCH	multicast channel
MCS	modulation and coding scheme
MDT	minimization of drive tests

MIB master information block
MIMO multiple-input multiple-output
MLD maximum likelihood detection
MME mobility management entity
MMEC MME code
MMSE minimum mean-square error
MRC maximal ratio combining
MSA MCH subframe allocation
MSI MCH scheduling information
MSP MCH scheduling period
MTC machine-type communications
MTCH multicast traffic channel
M-TMSI M-temporary mobile subscriber identity
MU-MIMO multi-user MIMO
NACK negative acknowledgment
NAS non-access stratum
NCC next-hop chaining counter
NDS network domain security
NH next hop key
NNSF NAS node selection function
NR neighbor cell relation
NRT neighbor relation table
OCC orthogonal cover code
OFDM orthogonal frequency division multiplexing
OFDMA orthogonal frequency division multiple access
OMC-ID operation and maintenance center identity
OOB out-of-band (emission)
OTDOA observed time difference of arrival (positioning method)
PA power amplifier
PAPR peak-to-average power ratio
PBCH physical broadcast channel
PBR prioritized bit rate
PCC primary component carrier
PCCH paging control channel
PCell primary cell
PCFICH physical control format indicator channel
PCH paging channel
PCI physical cell identifier
PCRF policy and charging rules function
PDCCH physical downlink control channel
PDCP packet data convergence protocol
PDN packet data network
PDSCH physical downlink shared channel
PDU protocol data unit

P-GW	PDN gateway
PHICH	physical hybrid ARQ indicator channel
PHR	power headroom report
PHY	physical layer
PLMN	public land mobile network
PMCH	physical multicast channel
PMI	precoding matrix indicator
PMI	precoding matrix index
PMIP	proxy mobile IP
PO	paging occasion
PRACH	physical random access channel
PRB	physical resource block
P-RNTI	paging RNTI
PRS	positioning reference signal
PSAP	public safety answering point
PSC	packet scheduling
pTAG	primary timing advance group
PTI	procedure transaction identifier
PUCCH	physical uplink control channel
PUSCH	physical uplink shared channel
PWS	public warning system
QAM	quadrature amplitude modulation
QCI	QoS class identifier
QoS	quality of service
QPP	quadrature permutation polynomial
RAC	radio admission control
RACH	random access channel
RAR	random access response
RA-RNTI	random access RNTI
RAT	radio access technology
RB	radio bearer
RBC	radio bearer control
RBG	resource block group
RE	resource element
REG	resource-element group
RET	remote electrical tilting
RF	radio frequency
RFSP	RAT/frequency selection priority
RI	rank indication
RIM	RAN information management
RLC	radio link control
RLF	radio link failure
RM	rate matching/Reed–Muller (code)
RN	relay node

RNC radio network controller
RNL radio network layer
RNTI radio network temporary identifier
ROHC robust header compression
RoT rise over thermal
R-PDCCH relay physical downlink control channel
RRC radio resource control
RRM radio resource management
RS reference signal
RSRP reference signal received power
RSRQ reference signal received quality
RSSI received signal strength indicator
RTP real-time transport protocol
RTT round trip time
RU resource unit
RV redundancy version
RX receiver
S1-MME S1 for the control-plane
S1-U S1 for the user-plane
SAE system architecture evolution
SAP service access point
SCC secondary component carrier
SCell secondary cell
SC-FDMA single carrier-frequency division multiple access
SCH synchronization channel
SCM spatial channel model
SDF service data flow
SDMA spatial division multiple access
SDU service data unit
SeGW security gateway
SFBC space-frequency block code
SFN system frame number
SGSN serving GPRS support node
S-GW serving gateway
SI system information
SIB system information block
SIC successive interference cancellation
SIMO single-input multiple-output
SINR signal to interference plus noise ratio
SIP session initiation protocol
SIPTO selected IP traffic offload
SI-RNTI system information RNTI
SISO single-input single-output/soft-input soft-output
SMS short message service

SN	sequence number
SNR	signal-to-noise ratio
SON	self-organizing network
SPID	subscriber profile ID for RAT/frequency priority
SPS C-RNTI	semi-persistent scheduling C-RNTI
SR	scheduling request
SRB	signaling radio bearer
SRS	sounding reference signal
SRVCC	single radio voice call continuity
sTAG	secondary timing advance group
S-TMSI	S-temporary mobile subscriber identity
SU	scheduling unit
SU-MIMO	single-user MIMO
TA	tracking area
TAC	tracking area code
TAD	traffic aggregate description
TAG	timing advance group
TAI	tracking area identity
TAU	tracking area update
TB	transport block
TCP	transmission control protocol
TDD	time division duplex
TDM	time division multiplexing
TEID	tunnel endpoint identifier
TFT	traffic flow template
TI	transaction identifier
TIN	temporary identity used in next update
TM	transparent mode
TMA	tower mounted amplifier
TNL	transport network layer
TPC	transmit power control
TPMI	transmitted precoding matrix indicator
TTI	transmission time interval
TX	transmitter
UCI	uplink control information
UDP	user datagram protocol
UE	user equipment
UL	uplink
UL TFT	uplink traffic flow template
ULR-Flags	update location request flags
UL-SCH	uplink shared channel
UM	unacknowledged mode
UMTS	universal mobile telecommunication system
UpPTS	uplink pilot time slot

UTRA	universal terrestrial radio access
UTRAN	universal terrestrial radio access network
VoIP	voice over Internet protocol
VoLTE	voice over LTE
VRB	virtual resource block
X2-C	X2-control plane
X2-U	X2-user plane
ZC	Zadoff–Chu
ZCZ	zero correlation zone
ZF	zero forcing (receiver)

This page intentionally left blank

CHAPTER

Introduction to LTE-Advanced

1

CHAPTER CONTENTS

1.1 Background on IMT-Advanced standards development 5
1.2 Requirements for the IMT-Advanced systems 8
1.3 Introduction to 3GPP standards 11
1.4 Evolution of 3GPP standards 18
1.5 A perspective on the future trends in cellular networks 20
References 26

Wireless communication comprises a wide range of technologies, services, and applications that have come into existence to meet the particular needs of users in different deployment scenarios. Wireless systems can be broadly characterized by content and services offered, reliability and performance, operational frequency bands, standards defining those systems, data rates supported, bidirectional and unidirectional delivery mechanisms, degree of mobility, regulatory requirements, complexity, and cost. The number of mobile subscribers has increased dramatically worldwide in the past two decades. The growth in the number of mobile subscribers will be further intensified by adoption of broadband mobile access technologies in largely populated and developing countries. It is envisioned that potentially the entire world population will have access to broadband mobile services depending on economic conditions and favorable cost structures offered by regional network operators. There are already more mobile devices than fixed-line telephones or fixed computing platforms such as desktop computers that can access the Internet. The number of mobile devices is expected to continue to grow more rapidly than nomadic and stationary devices. Mobile terminals will be the most commonly used platforms for accessing and exchanging information. In particular, users will expect a dynamic, continuing stream of new applications, capabilities, and services that are ubiquitous and available across a range of devices using a single subscription and a single identity. Versatile communication systems offering customized and ubiquitous services based on diverse individual

LTE-Advanced.
© 2014 Elsevier Inc. All rights reserved.

Samsung Ex. 1005, Page 32 of 1147

needs require flexibility in the technology in order to satisfy multiple demands simultaneously. Wireless multimedia traffic is increasing far more rapidly than voice and will increasingly dominate traffic flows. The paradigm shift from predominantly circuit-switched air-interface design to full IP-based systems has provided users with the ability to more efficiently, more reliably, and more securely utilize multimedia services including e-mail, file transfers, messaging, browsing, gaming, voice over Internet protocol (VoIP), and location-based multicast and broadcast services. These services can be either symmetrical or asymmetrical (in terms of use of radio resources in the downlink or uplink) and real-time or non-real-time with differing quality of service (QoS) requirements. The new applications consume relatively larger bandwidths, resulting in higher data rate requirements [13,14].

International Mobile Telecommunications-Advanced (IMT-Advanced) systems are broadband mobile wireless access systems that include new capabilities and flexibility that go beyond those of legacy IMT-2000 systems [1–7]. IMT-Advanced provided a global framework for development of the new generation of mobile services that offer fast data access and unified messaging, and broadband wireless multimedia in the form of new interactive services anywhere and anytime. Such systems provide access to a wide range of telecommunication services including advanced mobile services and are supported by mobile and fixed networks, which are entirely based on Internet protocol (IP). The IMT-Advanced systems support low to high mobility applications and a wide range of data rates in accordance with user and service demands in multiple user environments. The IMT-Advanced systems have also capabilities for high-quality multimedia applications within a wide range of services and platforms, providing a significant improvement in performance and QoS.

The capabilities of IMT-2000 systems have been continuously enhanced over the past decade as IMT-2000 technologies are upgraded and deployed. From the radio access perspective, the evolved IMT-2000 systems have been built on the existing systems and further enhanced the radio interface functionalities, and at the same time, new systems have emerged to replace the legacy radio access systems. This evolution has improved the reliability and throughput of the cellular systems and has promoted the development of an increasing number of services and applications. Similarity of services and applications across different IMT technologies and frequency bands is beneficial to users, and a broadly similar user experience leads to a large-scale deployment of products and services. The technologies, applications, and services associated with systems beyond IMT-Advanced could well be radically different from the present, challenging the perceptions of what may be considered viable by today's standards and going beyond what has just been achieved by the IMT-Advanced radio systems.

3GPP initiated a project on the long-term evolution (LTE) of the Universal Mobile Telecommunication System (UMTS) radio access network (RAN) in late 2004 to maintain 3GPP's competitive edge over other cellular technologies. The evolved UMTS terrestrial radio access network (E-UTRAN) substantially

improved end-user throughputs and system capacity and significantly reduced user-plane and control-plane latencies, bringing in considerably improved user experience with full mobility. With the emergence of IP as the protocol of choice for carrying all types of traffic, the 3rd Generation Partnership Project (3GPP) LTE has provided support for IP-based traffic with end-to-end QoS. Voice traffic is supported mainly as VoIP, enabling better integration with other multimedia services. Unlike its predecessors, which were developed within the framework of a global system for mobile communication (GSM)/UMTS network architecture, 3GPP specified an evolved packet core (EPC) architecture to support the E-UTRAN through reduction in the number of network elements and simplification of functionality, but most importantly allowing for connections and handover to other fixed and wireless access technologies, providing the network operators with the ability to deliver a seamless mobility experience. 3GPP set aggressive performance requirements for LTE that relied on improved physical layer technologies such as orthogonal frequency division multiplexing (OFDM) and single-user and/or multi-user multiple-input multiple-output (MIMO) techniques, and streamlined layer 2/layer 3 protocols and functionalities. The main objectives of LTE were to minimize the system and user equipment (UE) complexities, to allow flexible spectrum allocation in the existing or new frequency bands, and to enable coexistence with other 3GPP and non-3GPP radio access technologies. The LTE Rel-8/9 was used as the baseline for further enhancements under 3GPP Rel-10 to meet the requirements of the IMT-Advanced. A candidate proposal based on the latter enhancements (3GPP LTE-Advanced) was submitted to the International Telecommunication Union-Radio Telecommunications (ITU-R) and was subsequently qualified as an IMT-Advanced technology in 2010 [6,7]. However, concurrent with 3GPP LTE standard development, the operators were rolling out high-speed packet access (HSPA) networks to upgrade their second-generation (2G) and early 3G infrastructure; thus, they were not ready to embrace yet another paradigm shift in the RAN and core network technologies. Therefore, 3GPP continued to improve UMTS technology by adding multi-antenna support at the base station, higher modulation order in the downlink, and multi-carrier support to prolong the life span of 3G systems [15,16].

LTE-Advanced (LTE Rel-10/11) extended the capabilities of LTE Rel-8/9 by introducing new features such as carrier aggregation, which is the most significant and complex improvement provided by LTE-Advanced that can combine up to five 20 MHz radio frequency (RF) carriers to create a virtually wide bandwidth; enhanced multi-antenna techniques in the uplink and downlink (e.g., SU-MIMO and MU-MIMO, reference signals, channel state information measurement/reporting); cell coverage and throughput improvement by means of relay nodes, which connect to relay-enabled eNBs; enhanced inter-cell interference coordination to improve performance of heterogeneous networks; and coordinated multi-point transmission/reception (also known as network MIMO in the literature), which is a set of techniques using different forms of MIMO and beamforming to improve the performance at cell edges using coordinated scheduling, and transmitters/

receivers and antennas that are not collocated to provide greater spatial diversity to improve link reliability and data rate as well as system capacity [17,18]. The framework for the LTE-Advanced air-interface technology was mostly defined by the use of wider bandwidths up to 100 MHz, contiguous/non-contiguous spectrum deployments, also referred to as spectrum aggregation, and a need for flexible spectrum usage. In general, OFDM provides a simple means to increase bandwidth by adding additional subcarriers. Due to the discontinuous nature of the spectrum reserved for IMT-Advanced, the available bandwidth is fragmented. Therefore, the user terminals should be able to filter, process, and decode such a large variable bandwidth. The increased decoding complexity is one of the major challenges of the wider bandwidths [9].

Some early deployments of LTE-Advanced are expected in late 2013 with increasing adoption in 2014 and beyond [13,14]. LTE-Advanced is forward and backward compatible with LTE Rel-8/9, meaning that LTE Rel-8/9 devices will work on LTE-Advanced networks and LTE-Advanced capable terminals will work on the legacy LTE networks. The LTE-Advanced features offer performance gains and may considerably affect the system complexity and cost. For instance, higher order MIMO schemes up to 8×8 in the downlink will improve peak data rates and spectral efficiency; however, this feature also impacts the network side and the complexity of the UE, as the implementation of the antennas differs and additional baseband transmission/reception chains are needed. Moreover, while band aggregation does not affect spectral efficiency (since it is independent of bandwidth), cell-edge user throughput (due to insufficient signal to interference plus noise ratio (SINR) value, secondary cells cannot be utilized), and coverage (since coverage is determined by the link budget of the primary cell), it does increase the cost of network deployment. It is worth mentioning that the peak data rate is improved depending on the number of the aggregated carriers, further affecting the complexity and cost of the UE implementation.

This book begins with an introduction to the LTE-Advanced/IMT-Advanced systems and an overview of the 3GPP standards evolution and standardization process. The unique, top-down systems approach taken in this book necessitates beginning with the review of the overall network architecture and later examination of each and every significant network element in the evolved packet system (EPS) in Chapter 2. Once the access network and core network aspects of the LTE/LTE-Advanced systems are described, we turn our attention to the reference model and protocol structure of LTE/LTE-Advanced and discuss the operation and behavior of each entity (base station, mobile station, femto-cells, and relay stations) as well as functional components and their interactions, logical interfaces, protocol stack, and protocol terminations in Chapter 3. The remaining chapters of this book are organized consistently with the order of the protocol layers on the user-plane and control-plane, starting from the network layer and moving down to the physical layer. Each protocol layer has been described in a separate chapter in order for the readers to clearly understand the functions and operation of each layer. The overall operation of the mobile station and the base station and their corresponding state

machines are described in Chapter 4. Perhaps, this chapter is the most important part of the book as far as understanding the general operation of the system is concerned. Chapter 5 describes the radio resource control functions and protocols. The packet data convergence protocol, radio link control, and medium access control sublayers are described in Chapters 6, 7, and 8, respectively. The description of downlink and uplink physical layer algorithms and protocols has been divided into two separate parts (Chapters 9 and 10) for further clarity and structure. There are many commonalities between LTE Rel-8, 9, 10, and 11; however, release-specific features/functionalities, algorithms, or protocols which have been introduced in subsequent releases are clearly distinguished in the content so that they are not confused with the legacy components throughout the book. In this book, the term E-UTRA or LTE refers to both LTE and LTE-Advanced radio access technologies everywhere except where the feature or functionality does not exist or is not supported in the previous releases. The performance evaluation of LTE-Advanced against IMT-Advanced requirements is described in Chapter 11 where the performance metrics are defined and link-level and system-level simulation methodologies and parameters are elaborated. The theory and operation of LTE Rel-11 coordinated multi-point transmission/reception and LTE Rel-10/11 carrier aggregation schemes are described in Chapters 12 and 13. Multi-radio coexistence and enhanced inter-cell interference coordination techniques used in LTE Rel-10/11 are described in Chapter 14. In Chapter 15, we describe the positioning methods and evolved multimedia broadcast/multicast services. Each chapter contains a theoretical description of the features/functions and scrutinizes the real-life and practical implementation and deployment issues and remedies.

1.1 Background on IMT-Advanced standards development

The past two decades have witnessed a rapid growth in the number of subscribers and incredible advancement in the technology of cellular communication from simple, all-circuit-switched, analog first-generation systems with limited voice service capabilities, limited mobility, and small capacity to the fourth-generation (4G) systems with significantly increased capacity, advanced all-digital packet-switched all-IP implementations that offer a variety of multimedia services. With the increasing demand for high-quality wireless multimedia services, the radio access technologies continue to advance with faster pace toward the next generation of systems. The general characteristics envisioned for the fourth generation of the cellular systems included all-IP core networks, support for a wide range of user mobility, significantly improved user throughput and system capacity, reliability and robustness, seamless connectivity, reduced access latencies, etc.

IMT-Advanced, or alternatively 4G, cellular systems are mobile systems that extend and improve upon the capabilities of legacy IMT-2000 standards. The IMT-Advanced systems provide the users with access to a variety of advanced IP-based services and applications. The IMT-Advanced systems can support a wide range of data rates, with different QoS requirements, proportional to user

mobility conditions in multi-user environments. The key features of IMT-Advanced systems can be summarized as follows [1]:

- Enhanced cell and peak spectral efficiencies, and cell-edge user throughput to support advanced services and applications
- Lower airlink access and signaling latencies to support delay-sensitive applications
- Support of higher user mobility while maintaining session connectivity
- Efficient utilization of spectrum
- Inter-technology interoperability, allowing worldwide roaming capability
- Enhanced air-interface-agnostic applications and services
- Lower system complexity and implementation cost
- Convergence of fixed and mobile networks
- Capability of interworking with other radio access systems

These features enable IMT-Advanced systems to accommodate emerging applications and services. The capabilities of IMT systems have been continuously enhanced proportional to user demand and technology advancements in the past decade. However, the framework and overall objectives of the systems beyond IMT-Advanced are expected to be a paradigm shift in the design and development of radio interface, network architecture, and topology [1]. Present mobile communication systems have evolved by incremental addition of capabilities and enhancement of features of the baseline (legacy) systems. Examples include the evolution of UMTS family of standards in 3GPP. The systems beyond IMT-Advanced will be realized by functional fusion of the existing IMT systems, nomadic wireless access systems, and other wireless systems with high commonality and seamless interworking, in conjunction with the development of new and enhanced features.

Translating service/application requirements into requirements that are directly related to data transport over a wireless network commonly leads to considering a limited number of QoS attributes, such as user throughput, packet delay and/or delay variations (often referred to as delay jitter), bit/packet error rate, and other similar aspects. This motivates the introduction of service classes that group together services that are similar in terms of their requirements toward a network. Figure 1.1 shows four service classes (conversational, interactive, streaming, and background services) and their characteristics in terms of reliability, bit rate, and latency that were considered for the IMT-Advanced systems. We will further discuss these requirements and characteristics in the next sections. In Figure 1.1, BER denotes bit error rate which is a measure of reliability of communication link and is the ratio of the number of incorrectly received information bits to the total number of information bits sent within a certain time interval [2].

Prominent standards developing organizations such as 3GPP[1], and Institute of Electrical and Electronics Engineers Standards Association (IEEE-SA)[2] have

[1] http://www.3gpp.org

[2] http://standards.ieee.org

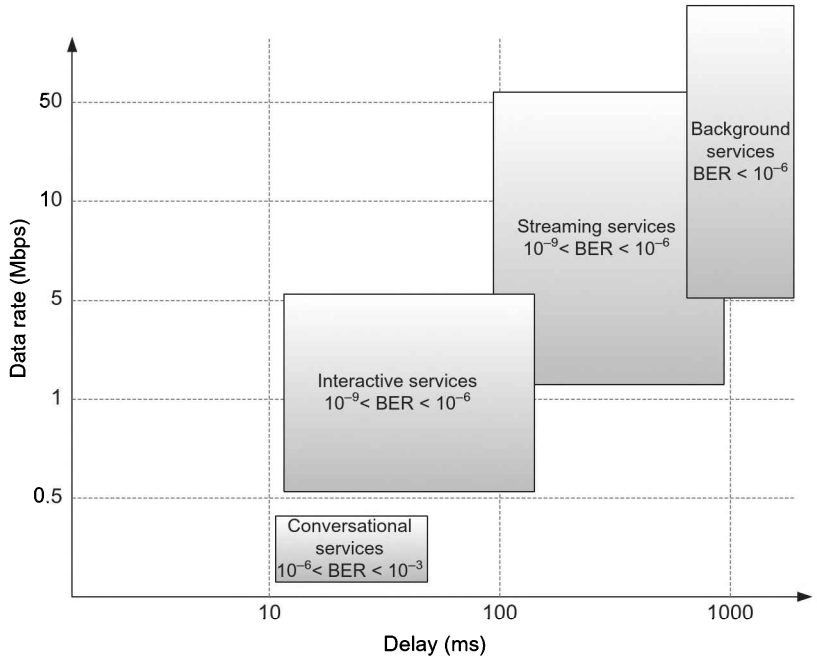


FIGURE 1.1

IMT-Advanced service classes and their characteristics [2].

actively contributed to the design, development, and proliferation of broadband wireless systems in the past decade. A number of broadband wireless access standards for fixed, nomadic, or mobile systems have been developed by these standardization groups and deployed by a large number of operators across the globe [6,7].

As mentioned earlier, the IMT-Advanced framework divided service classes into conversational, interactive, streaming, and background services. There are three service classes in the conversational service category. The basic conversational service class comprises basic services that are dominated by voice communication characteristics. The rich conversational service class consists of services that mainly provide synchronous communication enhanced by additional media such as video and collaborative document viewing. Conversational low-delay class covers real-time services that have very strict delay and delay variation requirements. In the interactive service category, two service classes are distinguished. Interactive services that permit relatively high delay which usually follow a request-response pattern (e.g., web browsing, database query). In such cases, response times in the order of a few seconds are permitted. Interactive services requiring significantly lower delay are remote server access or remote collaboration. In the streaming category, there are two service classes. The differentiating factor between these two classes is the live or non-live nature of

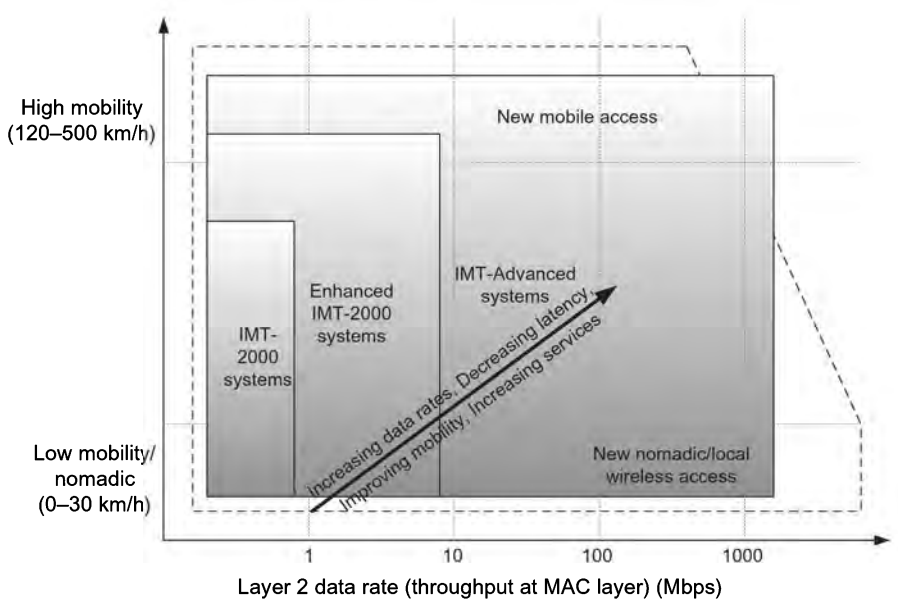


FIGURE 1.2

Illustration of the capabilities and evolution of IMT systems [1].

the content transmitted. In case of live content, buffering possibilities are very limited, which makes the service delay sensitive. In the case of non-live (i.e., pre-recorded) content, playout buffers at the receiver side provide a high robustness against delay and delay jitter. The background class only contains delay-insensitive services so that there is no need for further differentiation.

The advancement of capabilities and evolution of IMT systems are illustrated in Figure 1.2. As it appears from the figure, the services and performance of the systems noticeably increase as the systems evolve from one generation to another.

1.2 Requirements for the IMT-Advanced systems

The service and application requirements must be translated to system design requirements in order to obtain a set of minimum requirements and objective criteria for the design of a new mobile broadband wireless system. The design requirements encompass a wide range of system attributes such as data and signaling transmission latency over the airlink and the core network, system data rates, user cell-edge and average throughputs, and application capacity. The high-level requirements presented in Table 1.1 were defined by ITU-R Working Party 5D, prior to the development of the candidate proposals, for the purpose of consistent definition, specification, and evaluation of the IMT-Advanced systems.

Table 1.1 IMT-Advanced Requirements [3,5]

Cell Spectral Efficiency		
Test Environment	**Downlink (Bits/s/Hz/Cell)**	**Uplink (Bits/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
Peak Spectral Efficiency		
Test Environment	**Downlink (Bits/s/Hz)**	**Uplink (Bits/s/Hz)**
N/A	15	6.75
Cell-Edge User Spectral Efficiency		
Test Environment	**Downlink (Bits/s/Hz)**	**Uplink (Bits/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
Traffic Channel Link Data Rates		
Test Environment	**Uplink (Bits/s/Hz)**	**UE Speed (km/h)**
Indoor	1.0	10
Microcellular	0.75	30
Base coverage urban	0.55	120
High speed	0.25	350
VoIP Capacity		
Test Environment	**VoIP Capacity (Active Users/Sector/MHz)**	
Indoor	50	
Microcellular	40	
Base coverage urban	40	
High speed	30	

The intent of these requirements was to ensure that IMT-Advanced technologies were able to fulfill the objectives of the fourth generation of cellular systems, as laid out in Recommendation ITU-R M.1645 as the ITU-R vision for IMT-Advanced systems, as well as to set a specific level of minimum performance that each proposed technology needed to achieve [1]. These requirements were not

intended to restrict the full range of capabilities or performance that candidate technologies for IMT-Advanced might achieve, nor were they intended to describe how the IMT-Advanced technologies might perform in actual deployments under operating conditions that could be different from those presented in ITU-R recommendations and reports on IMT-Advanced. Satisfaction of these requirements were verified through link and system-level simulations (for VoIP capacity, spectral efficiencies, and mobility), analytical (for user and control-plane latencies as well as handover interruption time), and inspection (for service requirements and bandwidth scalability) [3]. The methodology, guidelines, and common configuration parameters were specified in Report ITU-R M.2135-1 [4].

The requirements for system and user data rates were described in the form of spectral efficiency to make them independent of bandwidth and practical spectrum allocations. Some of the requirements noted in Table 1.1 were evaluated via system-level simulations; however, the mobility requirement was evaluated through link-level simulation. Note that in Table 1.1, a cell or equivalently a sector is defined as a physical partition of the coverage area of a base station. The base station coverage area is typically partitioned into three or more sectors/cells to mitigate interference effects.

In Table 1.1, the cell spectral efficiency is defined as the aggregate throughput, i.e., the number of correctly received bits delivered at the data link layer over a certain period of time, of all users divided by the product of the effective bandwidth, the frequency reuse factor, and the number of cells. The cell spectral efficiency is measured in bits/s/Hz/cell.

The peak spectral efficiency in the above table denotes the highest theoretical data rate normalized by bandwidth (assuming error-free conditions) assignable to a single mobile station when all available radio resources for the corresponding link are utilized, excluding radio resources that are used for physical layer synchronization, reference signals, guard bands, and guard times (collectively known as layer-1 overhead). Bandwidth scalability is the ability to operate with different bandwidth allocations using single or multiple RF carriers. These values were defined assuming an antenna configuration of 4×4 and 2×4 in the downlink and uplink, respectively [3].

The normalized user throughput in Table 1.1 is defined as the number of correctly received bits by a user at the data link layer over a certain period of time, divided by the total spectrum. The cell-edge user spectral efficiency is defined as 5% point of cumulative distribution function (CDF) of the normalized user throughput.

Control-plane latency is defined as the transition time from idle state to connected state. The transition time (assuming downlink paging latency and core network signaling delay are excluded) of less than 100 ms is required for IMT-Advanced systems. The user-plane latency, also known as transport delay, is defined as the one-way transit time between a packet being available at the IP layer of the origin (user terminal in the uplink or base station in the downlink) and the availability of this packet at IP layer of the destination (base station in the uplink or

user terminal in the downlink). User-plane packet delay includes delay introduced by associated protocols and signaling assuming the user terminal is in the active mode. The IMT-Advanced systems are required to achieve a transport delay of less than 10 ms in unloaded conditions (i.e., a single user with a single data stream) for small IP packets (e.g., an empty payload + IP header) for both downlink and uplink as hypothetically shown in Figure 1.3. In Figure 1.3, the user-plane latency is measured from point 1 to point 2 or from point 3 to point 4 [3,19].

The following mobility classes were defined in IMT-Advanced systems [3]:

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

A mobility class is supported if the traffic channel link-level data rate, normalized by bandwidth, on the uplink, exceeds the minimum requirements shown in Table 1.1, when the user is moving at the maximum speed in that mobility class in each of the test environments. The requirements were defined assuming an antenna configuration of 4×2 and 2×4 in the downlink and uplink, respectively.

The handover interruption time is defined as the time duration during which a user terminal cannot exchange user-plane packets with any base station. The handover interruption time includes the time required to execute any RAN procedure, radio resource signaling protocol, or other message exchanges between the user terminal and the RAN. For the purposes of determining handover interruption time, interactions with the core network are not considered. It is also assumed that all necessary attributes of the target base station are known at initiation of the handover from the serving base station. The requirements for intra-frequency and inter-frequency handover interruption time were 27.5 and 40 ms (within a spectrum band)/60 ms (between spectrum bands), respectively [3].

In Table 1.1, VoIP capacity is calculated assuming a 12.2 kbps codec with 50% voice activity factor such that the percentage of users in outage is less than 2% where a user is defined to have experienced a voice outage if less than 98% of the VoIP packets have been delivered successfully to the user within a one-way radio access delay bound of 50 ms. The VoIP capacity is the minimum of the calculated capacity for downlink or uplink divided by the effective bandwidth in the respective link direction. These values were defined assuming an antenna configuration of 4×2 in the downlink and 2×4 in the uplink.

1.3 Introduction to 3GPP standards

The 3GPP unites several telecommunications standard development organizations (ARIB, ATIS, CCSA, ETSI, TTA, TTC), also known as organizational partners, and provides their members with a stable technical basis to transpose into their

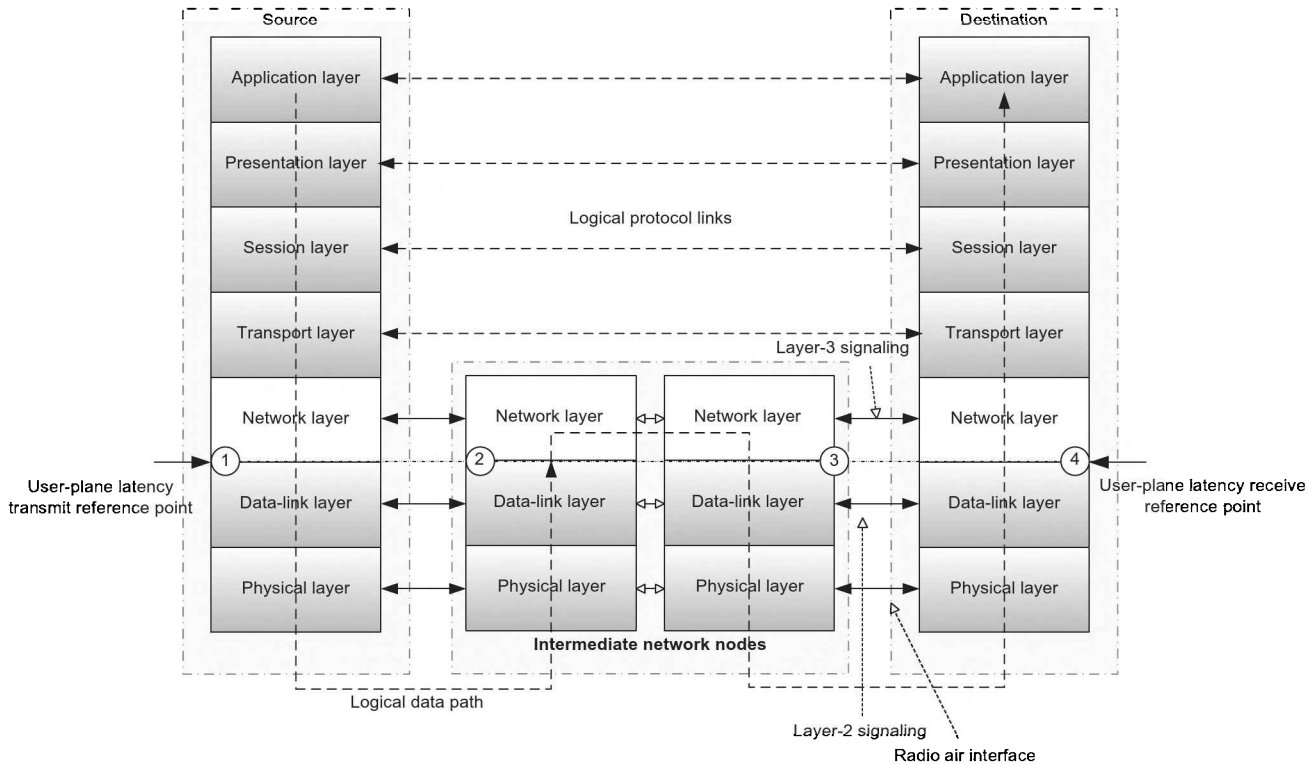


FIGURE 1.3

Graphical interpretation of user-plane latency [19].

own specifications. There are four technical specification groups in 3GPP which include RAN, Service and Systems Aspects (SA), Core Network and Terminals (CT), and GSM EDGE Radio Access Network (GERAN). Each technical group has a set of working groups and has a particular area of responsibility for the reports and specifications under its own terms of reference or group charter.

The original scope of 3GPP was to produce technical specifications and technical reports for 3G mobile systems based on evolved GSM core networks and the radio access technologies that they support, i.e., universal terrestrial radio access (UTRA) with both frequency division duplex (FDD) and time division duplex (TDD) modes. The scope was subsequently amended to include the maintenance and development of GSM technical specifications and technical reports including evolved radio access technologies, e.g., general packet radio service (GPRS) and enhanced data rates for GSM evolution (EDGE). The 3GPP organization was created in December 1998 by approval of the 3rd Generation Partnership Project Agreement. The latest 3GPP scope and objectives document has evolved from this original agreement. Table 1.2 shows the 3GPP organization and the charter of the technical specification groups and their subcommittees.

3GPP uses a system of parallel releases in order to provide device and network developers with a stable platform for implementation and to allow for addition of new features required by the market. Following the introduction of an entirely new air-interface with LTE in 3GPP Rel-8, and with various enhancements thereafter, the upcoming 3GPP Rel-12 and beyond will be as significant as ever in the industry's quest to extend mobile broadband availability, to provide more consistent service quality, and to economically satisfy the demand for fast and reliable mobile Internet, despite the challenges due to spectrum scarcity and other constraints. Successive mobile generations are typically depicted in technological terms. For the world's mobile operators following the development standards path, first laid by ETSI and then 3GPP, first-generation (1G) analog was followed by 2G TDMA-based GSM, then 3G wideband-CDMA-based UMTS, and 4G OFDM-based LTE/LTE-Advanced. However, concurrent with each upgrade in generation have been profound advances in the capabilities, economics, and availability of network, device, and service offerings.

3GPP specified support for several radio access technologies and also the handover between these access networks. The idea was to bring convergence using a unique core network providing various IP-based services over multiple access technologies. In addition to the support of the existing 3GPP radio access networks, 3GPP defined the interworking functions between the E-UTRAN (LTE/LTE-Advanced), GERAN (GSM/GPRS), and UTRAN (UMTS-based technologies WCDMA and HSPA). The EPS also allows the UE (and EPC) to access non-3GPP technologies. Non-3GPP accesses can be divided into two categories: the trusted and the untrusted. The trusted non-3GPP access networks can interact directly with the EPC, whereas the untrusted non-3GPP access network connects to the EPC via an intermediate interworking function and the evolved packet data

Table 1.2 3GPP Organization

Project Coordination Group (PCG) This group is the ultimate decision-making body in 3GPP responsible for final adoption of 3GPP Technical Specification Group work items, to ratify election results, and the resources committed to 3GPP	**GSM/EDGE Radio Access Network (GERAN)** This group is responsible for specifying the radio access part of GSM/EDGE, more specifically RF front-end, layer 1, 2, and 3, internal and external interfaces, conformance test specifications for all aspects of GERAN base stations and terminals, and GERAN specifications for the nodes in the GERAN	**GERAN WG1** This group is responsible for RF aspects of GERAN, internal GERAN interface specifications, specifications for GERAN radio performance and RF system aspects, conformance test specifications for testing of all aspects of GERAN base stations, etc. **GERAN WG2** This group is responsible for protocol aspects of GERAN. It specifies the data link layer protocols and the interfaces between these layers and the physical layer **GERAN WG3** This group is responsible for conformance test specifications for testing of all aspects of GERAN terminals and liaising with other technical groups to ensure overall coordination
	Radio Access Network (RAN) This group is responsible for the definition of the functions, requirements, and interfaces of the UTRA/E-UTRA network in its two modes, FDD and TDD.	**RAN WG1** This group is responsible for the specification of the physical layer of the radio Interface for UE, UTRAN, E-UTRAN, covering both FDD and TDD modes of radio interface. It is also responsible for handling physical layer related UE capabilities and physical layer related parameters used in UE tests developed in the RAN working group **RAN WG2** This group is in charge of the radio interface architecture and protocols, the specification of the radio resource control protocol, radio resource

(*Continued*)

Table 1.2 (Continued)

		management, and the services provided by the physical layer to the upper layers
		RAN WG3 This group is responsible for the overall UTRAN/E-UTRAN architecture and the specification of inter-base station protocols and interface between mobile station and base station
		RAN WG4 This group works on the radio aspects of UTRAN/E-UTRAN. It conducts simulations of various practical radio system scenarios and derives the minimum requirements for transmission and reception parameters. Once these requirements are set, the group defines the test procedures that will be used to verify them (only for base stations)
		RAN WG5 This group specifies conformance testing at the radio interface for the UE. The test specifications are based on the requirements defined by other groups such as RAN WG4 for the radio test cases, and RAN WG2 and CT WG1 for the signaling and protocols test cases
	Service and System Aspects (SA) The SA working group is responsible for the overall architecture and service capabilities of systems based on 3GPP specifications and has a responsibility for inter-group coordination	**SA WG1 (Services)** This group works on the services and features for 3G/4G and beyond systems. The group sets high-level requirements for the overall system and provides this in a stage 1 description in the form of specifications and reports
		SA WG2 (Architecture) This group is in charge of

(*Continued*)

Table 1.2 (Continued)

		developing stage 2 of the 3GPP network and identifies the main functions and entities of the network, how these entities are linked to each other, and the information they exchange. The group has a system-wide view, and decides on how new functions integrate with the existing network entities
		SA WG3 (Security) This group is responsible for the security of the 3GPP system, performing analyses of potential security threats to the system, considering the new threats introduced by the IP-based services and systems, and setting the security requirements for the overall 3GPP system
		SA WG4 (Codec) This group develops specifications for speech, audio, video, and multimedia codecs, in both circuit-switched and packet-switched environments
		SA WG5 (Telecom Management) This group specifies the management framework and requirements for management of the 3G system, delivering the architecture descriptions of the telecommunication management network and coordinating across other working groups
	Core Network and Terminals (CT) This group is responsible for specifying terminal interfaces,	**CT WG1** This group is responsible for the 3GPP specifications that define the UE-core network

(*Continued*)

Table 1.2 (Continued)

	terminal capabilities, and the core network part of 3GPP systems	layer-3 radio protocols and core network side of the reference points
		CT WG3 This group specifies the bearer capabilities for circuit and packet-switched data services, and the necessary interworking functions towards both, the UE and the terminal equipment in the external network
		CT WG4 This group defines stage 2 and stage 3 aspects within the core network focusing on supplementary services, basic call processing, mobility management within the core network, bearer independent architecture, GPRS between network entities, transcoder free operation, etc.
		CT WG6 This group is responsible for development and maintenance of specifications and associated test specifications for the 3GPP smart card applications, and the interface with the mobile terminal

http://www.3gpp.org/Specification-Groups

gateway (ePDG) entity. The main role of the ePDG entity is to provide security mechanisms such as IPsec tunneling of connections with the UE over an untrusted non-3GPP access. 3GPP does not specify which non-3GPP technologies should be considered trusted or untrusted. This decision is made by the operator.

The 3GPP standardization process starts with developing a requirement document where the targets to be achieved are determined (i.e., Stage 1). The system

architecture is defined where the underlying building blocks and interfaces are decided (Stage 2). Various working groups are engaged in developing the detailed specifications where every functional component, interface, and signaling protocol is specified (Stage 3). The compliance and interoperability testing and verification follow upon completion of Stage 3 specifications.

1.4 Evolution of 3GPP standards

As mentioned earlier, the growing demand for mobile Internet and wireless multimedia applications has motivated development of broadband wireless access technologies in the recent years. As a result, 3GPP initiated the work on the LTE in late 2004 to maintain 3GPP's competitive edge over other cellular technologies. The 3GPP LTE symbolizes the migration of the UMTS family of standards from systems that supported both circuit-switched and packet-switched voice/data communications to an all-IP, packet-switched system. The development of the LTE air-interface was closely coupled with the 3GPP system architecture evolution (SAE) project to define the overall system architecture and evolved packet core network [10]. Figure 1.4 illustrates the evolution of the 3GPP standards in the past decade, their major features, and release dates. It shows that the 3GPP standards have continuously evolved toward higher performance and data rates, lower access latencies, and increasing capability to support emerging wireless applications. To achieve higher downlink and uplink data rates, 3G operators upgraded their networks with high-speed downlink packet access (HSDPA), which was specified in 3GPP Rel-5, and high-speed uplink packet access (HSUPA), which was specified in 3GPP Rel-6. The HSDPA and HSUPA, collectively known as HSPA, were further upgraded in 3GPP Rel-7 and 8 with additional features such as higher order modulations in the downlink and uplink, downlink open-loop MIMO, improvements of data link layer protocols, and continuous connectivity for packet data users [13,14].

The 3GPP Rel-8 specifications specified LTE (access network) and EPC (core network) which were completed in late 2008 followed by the UE conformance testing specifications. The 3GPP then worked toward development of an advanced standard specification to address the IMT-Advanced requirements and services for the 4th generation of cellular systems [6,7]. The LTE-Advanced was part of 3GPP Rel-10 and was completed in early 2011. The IMT-Advanced compliant technologies may be available for commercial deployment as early as 2014. 3GPP Rel-11 which initially started as an incremental development, turned into a full release with several features such as coordinated multi-point transmission, enhanced carrier aggregation, and further enhanced inter-cell interference coordination. This release was completed in early 2013. 3GPP is currently working on Rel-12 whose main study and work items are described in Section 1.5.

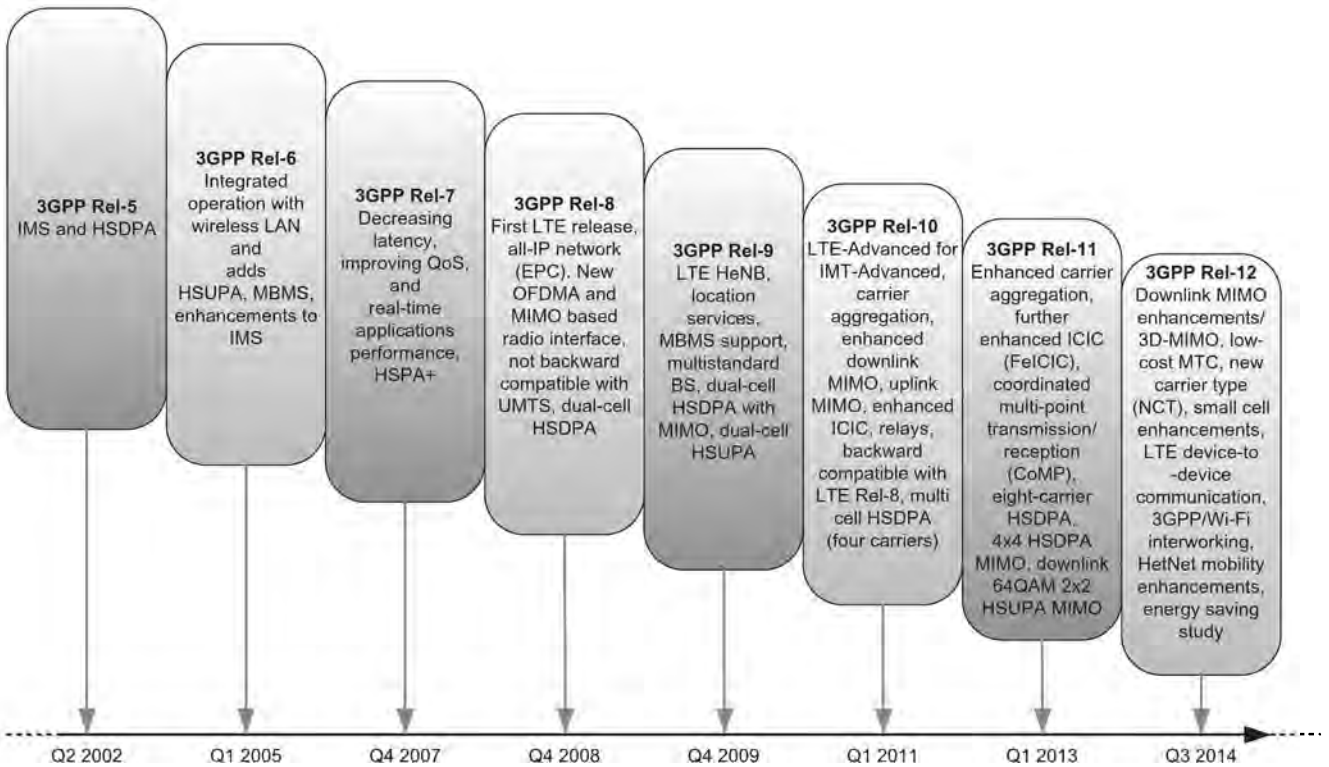


FIGURE 1.4

Evolution of the 3GPP standards in the past decade, their major features and release dates.

Source: *3GPP.*

1.5 A perspective on the future trends in cellular networks

A major challenge facing mobile operators and their technology suppliers is to satisfy the enormous user demands for mobile Internet applications and services. The number of users with mobile broadband wireless devices is exponentially increasing. The use of personal data services is growing with always on, always with you access to an expanding set of applications and services including network-intensive video streaming and cloud services. LTE networks are already providing peak data rates approaching 100 Mbps, but these are only possible under ideal conditions on lightly loaded networks and where UE is close to the base station radio transmitter [11].

While the existing frequency bands are often rearranged for more efficient use, new licensed bands, including higher frequencies for hot-spot zones are introduced, which will be used in conjunction with unlicensed spectrum (if suitable) along with cognitive radio techniques to access and manage the spectrum utilization. A large number of innovative techniques are being studied and researched which can increase spectral efficiency, reduce latency, and improve mobility and connectivity with emphasis on average achievable throughputs across the entire cell including cell-edges. Improving consistency of achievable performance across the network (capacity and throughput), rather than just increasing the peak data rates, is going to be a fundamental improvement demanded by the operators. The coordinated multi-point transmission/reception, and other interference coordination techniques will improve cell-edge performance. The 3D-MIMO and beamforming with large arrays of antenna elements would enable additional frequency reuse within a cell.

While the peak spectral efficiency continues to be important from a technology marketing perspective, the end-user experience is closely associated with the degree of uniformity of service provision; this implies that future enhancements will need to augment cell-edge data rates rather than simply addressing peak data rates. This means that system-level enhancements must generally be targeted considering multi-cell operation and coordination in order to realize useful improvements in spectral efficiency and user experience across the network. As the number of LTE/LTE-Advanced capable devices increases, smaller cell sizes are needed in order to deliver sufficient capacity; such capacity increases resulting from reducing the cell size are often referred to as cell-splitting gains. Since the small cells are embedded in the existing network of macro-cells, the result is a heterogeneous network of macro- and small cells operating at different transmission powers and with different coverage areas. The magnitude of the cell-splitting gains is dependent on suitable cell association/handover decisions to ensure that sufficient users are served by the small cells. In such networks, inter-cell interference coordination becomes significantly more complex than in a homogeneous network [12,16].

The new network topology comprising the small cells is seen as one of the key techniques to increase spectrum efficiency and to improve coverage, bringing the users closer to the network nodes (e.g., pico-cells, femto-cells, or remote radio

heads) as a result of the shrinking cell sizes. This would improve the SINR and increase the overall user throughput and capacity. The heterogeneous networks can impact many aspects of the system operation including large traffic and user variations in the cells, larger SINR dynamics, increased handover rates, and so on. One of the methods to deploy small cells overlaid by macro-cells is a frequency-separated local access, where different frequency layers are being used for the small cells. Another deployment scenario for small cells is a frequency-integrated local access, where macro- and pico-cells/femto-cells use the same frequency wherever the small cells are integrated into the network. Depending on which scenario is adopted, different methods are required to assign cell identities and to support the UEs in finding and using the respective small cells. For the small cells, particularly when using higher frequencies, new regulatory requirements might be required and coexistence studies need to be conducted. The use of different TDD schemes can also become beneficial. A more dynamic uplink/downlink subframe allocation in isolated cell clusters can more efficiently match to the instantaneous local traffic conditions. Advanced sensing and resource reservation might be required to avoid severe interference scenarios. The integrated local access is likely to improve performance. The small cells extend the macro-cell coverage with the same physical layer identification or with dynamically assigned virtual cell identities that are reused in different geographical areas. The small cells can provide a fast radio link with a largely improved link budget due to the close proximity of the network node to the user. The macro-cell, on the other hand, can provide cell-wide system information, as well as radio resource control including traffic management and carrier selection and supporting the UE to detect a neighboring small cell. The mobility in heterogeneous networks can be more robust and the operation of small cells may even be transparent to the UE, so the UE may not be aware of the local access. On the other hand, this architecture requires a tight integration of the small cells into the network by low-latency interfaces with a central baseband processing unit [12,16].

Improved backhaul and inter-eNB interfaces can support techniques such as baseband pooling and inter-cell coordination that can more efficiently and effectively manage resources with the large arrays of remote radio heads that will be deployed in high-demand locations. There are initiatives to simplify management across cells with simplified signaling and control maintained at the macrolayer, and introduction of phantom cells in high-density small-cell layers. Local wireless access overlaid by wide area macro-cells providing basic coverage has been identified as one of the key enablers of the future RAN. World Radiocommunication Conference 2015 (WRC-15) is expected to allocate the frequency range beyond 10 GHz to overcome spectrum shortage for broadband wireless communication systems, which may result in the development of a completely new access scheme potentially standardized as part of 3GPP future releases. Access technologies making extensive use of beamforming will be possible candidates since the beamforming gain may compensate the increased path loss at higher frequencies.

Although an increase in available and useful spectrum is expected to be at least threefold until 2020, this is by far not sufficient enough to keep up with the expected growth of traffic [20,21].

Downlink MIMO enhancements and three-dimensional MIMO is one of the 3GPP Rel-12 study items. With continued advancement of multi-antenna technologies, enhancements of the macro-cell eNBs can be realized mostly by exploiting an increased frequency reuse and reduced interference in the spatial domain. Under the new study item, a new three-dimensional channel model will be developed to allow more accurate system-level simulation of multi-antenna technologies. Three-dimensional beamforming through utilization of flexible electronic beam shaping adds the concept of beamforming in the vertical dimension. It has the potential to increase spectrum efficiency of the network through proactive cell shaping and splitting as well as improving the coverage. Other enhancement areas to support this technology may include reference signals, codebooks and feedback, and measurement and reporting for interference coordination in the beam domain as well as RF requirements.

The significant growth of machine-to-machine (M2M) applications (e.g., smart grids, transportation, healthcare services) has compelled the 3GPP and the cellular industry to ensure that the design and development of future releases of LTE can support massive transfer of small, infrequent packets using very low cost, low complexity, and low power devices. The 3GPP version of M2M, called machine type communication (MTC), was already standardized in Rel-11. The latter work covers service requirements, architecture, and security issues. An MTC interworking function and service-capability server, among other functionalities, were defined as part of 3GPP Rel-11 [12,22].

The LTE-Advanced carrier aggregation scheme in Rel-10/11 was limited to backward-compatible carriers only. On the one hand, this enables a smooth transition to new releases; on the other hand, it makes the introduction of new technologies more difficult and cumbersome. One of the consequences of backward compatibility to LTE Rel-8/9 is the necessity to continuously transmit cell-specific reference signals in every downlink subframe across the system bandwidth. This precludes switching off a cell temporarily and represents unnecessary overhead particularly in the case of non-codebook-based beamforming using UE-specific reference signals or multi-user MIMO. A new carrier type allows switching off cells, at least temporarily, and will reduce the overhead and interference from cell-specific reference signals by maximizing the use of UE-specific reference signals that can be precoded with the data subcarriers which are required for advanced multi-antenna technologies. Furthermore, new bandwidth combinations may be introduced to enable more efficient use of fragmented bandwidths. The new carrier type can be operated as an extension carrier along with another LTE/LTE-Advanced carrier, or alternatively as a stand-alone non-backward-compatible carrier in Rel-12 and beyond [16].

Device-to-device (D2D) communication is a new feature in LTE Rel-12 that allows direct communication between user terminals that are in the proximity of

each other. In addition to its potential for energy saving, interference reduction, and coverage extension, the key driver for this technology is to ensure that the evolution of LTE meets the growing demands for new applications/services. Once D2D functionality is developed, the market may find new proximity-based applications and services. Focus is mainly given to network-controlled D2D communication. In this case, the attach procedure, e.g., initial access, authentication, connection control, as well as resource reservation, can be handled by the network; therefore, the QoS can be guaranteed and the network operator can remain in control of the transactions (because they own the infrastructure and the spectrum).

Today's Wi-Fi network detection and selection functions are mostly UE-based, and future solutions might be operator controlled. This will empower implementation of more intelligent networks that monitor cell load or transport network load, user QoS requirements, or radio link quality for the different links during the operation. Cellular networks supporting Wi-Fi are seen as a viable solution by operators to off-load network traffic in indoor or dense environments, but until now, interworking has been mainly limited to core network functions such as user authentication/authorization and accounting. Although 3GPP specifications support mobility across 3GPP and non-3GPP access networks, further optimization is needed to improve load balancing, QoS provisioning, mobility and handover across heterogeneous networks, and improved UE power consumption when using Wi-Fi radio in tight cooperation with cellular networks [12,16].

The future evolution path or further enhancements of LTE-Advanced system will be directed toward a balance among determining factors such as technical solutions, practicality, and cost of implementation and deployment. Not every theoretical claim of potential improvement is deemed cost effective and practical to deploy. Factors such as robust conformance testing and real-life performance verification must also be taken into account, and unnecessary market fragmentation must be avoided, in order to ensure that economies of scale are achieved. Such factors would imply that the introduction of multiple alternative options, offering similar performance, should be avoided; careful evaluation must continue to make informed selection of new features [16].

It is evident that the existing mobile broadband radio access systems will continue to evolve and new systems will emerge. The vision, service and system requirements as well as use cases for systems beyond IMT-Advanced (alternatively known as 5G) are being defined. While it is not exactly clear what technologies will be incorporated into the design of such systems and whether the existing radio access technologies will converge into a single universal radio interface, it is envisioned that the future radio interfaces will rely on software-defined radio, cognitive radio system, flexible spectrum sharing, small-cell network deployments, cloud-based services, and cooperative communication concepts as well as reconfigurable RF and baseband circuitry in order to provide higher quality of user experience, higher capacities, and wider range of services with minimal cost and complexity.

Although the LTE family of standards have been designed from the beginning to facilitate energy-efficient implementations of both mobile devices and base stations, energy efficiency is an area with increasing importance due to the rising cost of energy and environmental concerns; thus, the evolution of LTE-Advanced must prioritize such considerations. This will involve minimizing the transmission overheads and processing requirements when network load is light, and in general enhancing the ability of the network to adapt dynamically to the actual load levels. This may include more advanced techniques for load balancing and resource sharing so that parts of the system which are not essential may be switched off whenever possible, while retaining flexibility to bring them online rapidly to handle high traffic demand when needed [16].

A generic framework for coordinated multi-point transmission/reception in the downlink and uplink was developed in LTE Rel-11, which may be deployed as part of LTE-Advanced networks. The gains promised by this method may only be achieved upon availability of low-latency broadband fiber-optic interconnections between coordinated transmission points. Such links are not widely available in contemporary networks; thus, early CoMP deployments should not be overly dependent on their existence. Nevertheless, centralized (alternatively known as cloud) processing may become an important enabler for realizing the full potential of CoMP, and efficient schemes for the transfer of scheduling/precoding information as well as data between transmission points will need to be identified. It is worth noting that centralized processing has the potential to provide some practical benefits in addition to those arising from centralized CoMP scheduling/beamforming and joint system optimization. The centralized processing can potentially reduce the costs of network operation by reducing the number of secure sites needed for network equipment installation, as well as facilitating network reconfiguration and robustness. For the uplink CoMP, each reception point can use interference cancellation techniques to isolate the weaker signals from mobile devices in the neighboring cells and convey them to a central node for combining with the versions of the signal received at other points.

MU-MIMO techniques are the most effective solutions to increase spectral efficiency in a cell. However, some of the promised MU-MIMO gains in practice fade away. The gap between theoretical MU-MIMO performance and that practically achievable in a real-life deployment may be explained by arguing that good spatial separation of users is achieved with correlated arrays of uniformly spaced antenna elements, whereas many antenna configurations deployed in practice are cross-polarized, with low correlation between the polarizations. Furthermore, channel state information at the transmitter is far from perfect, due to the constraints of uplink feedback overhead and latency. The availability of sufficiently spatially separated users with non-empty data queues is limited in practice; nevertheless, as the number of LTE-Advanced capable UEs increases, this will become less of a concern. The studies have shown that for a given channel state information feedback overhead, higher capacity can be achieved with more accurate channel state information feedback from a small number of terminals relative to

less accurate feedback from a large number of terminals when using MU-MIMO techniques. Therefore, future improvements in downlink MU-MIMO will depend on significant improvement in the accuracy of the channel state information at the transmitter, given practical channel estimation, realistic antenna configurations, typical mobility scenarios, without incurring unrealistic feedback overhead. In the uplink, MU-MIMO can be implemented without availability of accurate channel state information at the transmitter. The complexity of the uplink MU-MIMO operation is managed at the receiver rather than the transmitters, and sophisticated non-standardized receivers can theoretically attain the maximum multiple access channel capacity [16].

Another promising future trend in cellular networks is software-defined networking (SDN) or network virtualization. The SDN is an emerging network architecture where the control and data planes are decoupled, network intelligence and state are logically centralized, and the underlying network infrastructure is abstracted from the applications. The migration of control layer into accessible computing devices (formerly tightly coupled with individual network devices) enables the underlying infrastructure to be abstracted for applications and network services, which can treat the network as a logical or virtual entity. Figure 1.5 illustrates the concept of a software-defined network. In one embodiment of SDN, the network control-plane hardware can be physically decoupled from the data plane hardware; i.e., a network switch can forward packets and a separate server can run the network control plane. As a result, enterprises and carriers achieve programmability, automation, and network control, enabling them to build highly

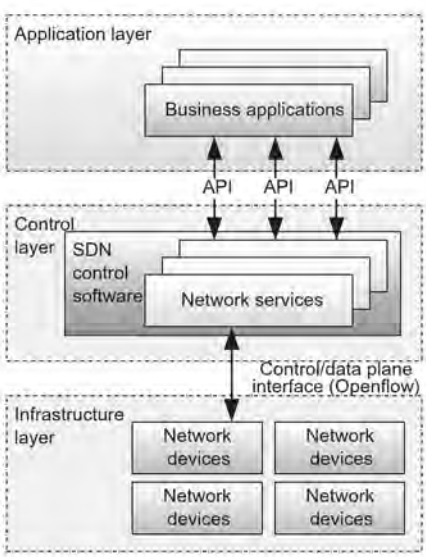


FIGURE 1.5

Illustration of the software-defined network concept [23,24].

scalable, flexible networks that readily adapt to changing business needs. The rationale for this approach is that the decoupling allows for the control plane to be implemented using a different distribution model than the data plane, and the decoupling further allows the control-plane development and run-time environment to be on a different platform than the traditionally low-power management CPUs found on hardware switches and routers [23,24].

The SDN requires some method for the control plane to communicate with the data plane. One example of such a mechanism is open flow, which is a standard interface for controlling computer network switches. The SDN allows network operators to have programmable central control of network traffic without requiring physical access to the network's hardware devices. In addition to abstracting the network, SDN architecture supports a set of application programming interfaces (APIs) that make it possible to implement common network services, including routing, multicast, security, access control, bandwidth management, traffic engineering, QoS, processor and storage optimization, energy usage, and all forms of policy management, which are customized to meet business objectives. The Open Networking Foundation is studying open APIs to promote multivendor management, which opens the door for on-demand resource allocation, self-service provisioning, virtualized networking, and secure cloud services. Using open APIs between the SDN control and the applications layers, the business applications can operate on an abstraction of the network, leveraging network services and capabilities without being tied to the details of their implementation [23,24].

References

ITU Recommendations and Reports

[1] Recommendation ITU-R M.1645, Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000, June 2003.

[2] Recommendation ITU-R M.1822, Framework for services supported by IMT, October 2007.

[3] Report ITU-R M.2134, Requirements related to technical performance for IMT-Advanced radio interface(s), October 2008.

[4] Report ITU-R M.2135-1, Guidelines for evaluation of radio interface technologies for IMT-Advanced, December 2009.

[5] Report ITU-R M.2133, Requirements, evaluation criteria and submission templates for the development of IMT-Advanced, October 2008.

[6] Recommendation ITU-R M.2012, Detailed specifications of the terrestrial radio interfaces of International Mobile Telecommunications Advanced (IMT-Advanced), January 2012.

[7] IMT-Advanced submission and evaluation process < http://www.itu.int/ITU-R/ > .

[8] ITU-T Next Generation Networks Global Standards Initiative, < http://www.itu.int/ITU-T/ngn/ > .

3GPP Specifications

[9] 3GPP TR 36.913, Requirements for further advancements for Evolved Universal Terrestrial Radio Access (E-UTRA) (LTE-Advanced) (Release 8), March 2009.
[10] 3GPP TR 23.882, 3GPP System Architecture Evolution: Report on Technical Options and Conclusions (Release 8), September 2008.

Articles

[11] K. Mallinson, 2020 Vision for LTE, WiseHarbor. < http://www.wiseharbor.com/forecast.html >, June 2012.
[12] E. Seidel, 3GPP LTE-Standardization in Release 12 and Beyond, Nomor Research GmbH, Munich, Germany. < http://www.nomor.de/lte-standardisation >, January 2013.
[13] 4G Americas, 4G Mobile Broadband Evolution: Release 10, Release 11 and Beyond - HSPA, SAE/LTE and LTE-Advanced. < http://www.4gamericas.org/ >, October 2012.
[14] 4G Americas, Mobile Broadband Explosion: The 3GPP Wireless Evolution. < http://www.4gamericas.org/ >, August 2012.
[15] P. Bhat, et al., LTE-Advanced: an operator perspective, IEEE Commun. Mag. (February 2012).
[16] M. Baker, From LTE-Advanced to the future, IEEE Commun. Mag. (February 2012).
[17] M. Kottkamp, A. Roessler, J. Schlienz, LTE-Advanced Technology Introduction, Rohde & Schwarz, August 2012.
[18] I.F. Akyildiz, et al., The evolution to 4G cellular systems: LTE-Advanced, Phys. Commun. 3 (4) (December 2010) < http://www.sciencedirect.com/ >
[19] White Paper, NGMN Radio Access Performance Evaluation Methodology, NGMN Alliance, January 2008.
[20] White Paper, LTE Backhauling Deployment Scenarios, NGMN Alliance, July 2011.
[21] White Paper, Guidelines for LTE Backhaul Traffic Estimation, NGMN Alliance, July 2011.
[22] White Paper, On Machine-to-Machine Communication, NGMN Alliance, September 2012.
[23] White Paper, Software-Defined Networking: The New Norm for Networks, Open Networking Foundation, April 2012. < https://www.opennetworking.org >.
[24] Software-defined networking, from Wikipedia, < http://en.wikipedia.org/wiki/Software-defined_networking >.

This page intentionally left blank

CHAPTER

Network Architecture 2

CHAPTER CONTENTS

2.1 Overall network architecture 31
2.1.1 General concepts 31
2.1.2 Evolution of the network architecture 38
2.2 EPC architecture 42
2.2.1 Reference model and network interfaces 42
2.2.2 Public land mobile network 43
2.2.3 Packet data network gateway 46
2.2.4 Serving gateway 47
2.2.5 Mobility management entity 47
2.2.6 Policy and charging rules function 48
2.2.7 Home subscriber server 49
2.2.8 Authorization, authentication, and accounting 49
2.2.9 EPC identifiers 50
2.3 Roaming architecture and interworking with other networks 50
2.4 E-UTRAN architecture 57
2.4.1 E-UTRAN functions 59
2.4.2 E-UTRAN interfaces 59
2.4.3 E-UTRAN identifiers 60
2.4.4 Support of HeNB in E-UTRAN 61
2.4.5 Support of RRHs in E-UTRAN 63
2.4.6 Support of relays in E-UTRAN 64
2.5 E-MBMS architecture 67
2.6 Positioning network architecture 71
2.7 Mobility management 74
2.8 Radio resource management 75
2.9 Quality of service/quality of experience 77
2.10 Security in E-UTRAN 78
2.11 Support of voice and messaging services in LTE/LTE-Advanced 80
2.11.1 IMS architecture 80
2.11.2 Single radio voice call continuity 84
2.11.3 Voice over LTE 90
2.11.4 Circuit-switched fallback 94

LTE-Advanced.

© 2014 Elsevier Inc. All rights reserved.

Samsung Ex. 1005, Page 60 of 1147

2.12 LIPA, selected IP traffic offload, and IP flow mobility 100
2.12.1 Local IP access 101
2.12.2 Selected IP traffic offload 105
2.12.3 IP flow mobility 107
2.12.4 Wi-Fi offloading 110
2.13 Mobile IP 111
2.13.1 Client-based IP mobility 113
2.13.2 Network-based IP mobility 113
2.14 Network architecture evolution and small cells 114
References 116

3GPP Rel-8 specifications included enhancements to high-speed packet access (HSPA) core and access networks as well as the introduction of the evolved packet system (EPS) which comprises an Internet protocol (IP)-based, all-packet core network with minimal hierarchy in conjunction with a new orthogonal frequency division multiplexing (OFDM)-based radio access network. The EPS is a simplified architecture (i.e., some intermediate network elements were removed or combined together) with an all-IP network support for higher throughput, lower latency radio access and improved mobility between multiple heterogeneous access networks including evolved universal terrestrial radio access (E-UTRA), 3GPP legacy systems (e.g., GERAN or UTRAN), as well as non-3GPP systems, such as Wi-Fi and CDMA2000.

The network architecture of LTE is based on the functional decomposition principle, where underlying features are divided into logical functional entities without specific implementation assumptions about physical network nodes. That is why 3GPP specified an entirely new packet-switched core network architecture to support the evolved universal terrestrial radio access network (E-UTRAN) through reduction in the number of network elements, simpler functionality, and more importantly allowing connections and handover to other access technologies, giving the service providers the ability to deliver a seamless mobility experience.

The introduction of the evolved packet core (EPC) and all-IP network architecture had significant impacts on mobile services, where voice, data, and video communications are built on the IP-based protocols; interworking of the new architectural entities with 2G/3G legacy systems; scalability required by each of the core elements to address the increasing number of direct connections to user terminals, orders of magnitude of bandwidth increase, and terminal mobility; reliability and availability provided by each network entity to ensure service continuity. To address a fundamentally different set of network and service requirements, the EPC design represented a departure from previous mobile networking paradigms.

This chapter describes the design principles of the EPS, network reference model, logical building blocks of the network architecture as well as theoretical

Samsung Ex. 1005, Page 61 of 1147

and operational aspects of EPC and LTE/LTE-Advanced radio access network (E-UTRAN) building blocks, functional split, and their interconnections. Under this topic, the network functions, such as mobility management, radio resource management (RRM), quality of service (QoS)/QoE, mobile IP, IP flow mobility, relays and femto-cells, and security, along with practical design/implementation considerations and real-life issues and limitations will be discussed. Furthermore, the operation and management of small cells involving low-power remote radio heads (RRHs) will be covered. The selective IP traffic offload, local IP access (LIPA), and Wi-Fi offloading mechanisms will be discussed, as well. In this chapter, we provide a top-down systems approach to understanding the architecture and operation of the LTE/LTE-Advanced access and core networks. The focus in this chapter is on the network architecture and underlying network functional entities and the network interfaces. The protocols and procedures for network attach/detach, connection establishment/release, mobility management, and other functional details will be discussed in the next chapters.

2.1 Overall network architecture

2.1.1 General concepts

In its most abstract form, the EPS consists of only two nodes in the user-plane, that is, a base station and a core network gateway. The node that performs control-plane functionality is separated from the node that performs user-plane functionality. The basic EPS architecture is illustrated in Figure 2.1. The EPS architecture was designed not only to provide a smooth evolution from the 2G/3G quasi-packet-switched architectures consisting of nodeBs (NBs), radio network controllers (RNCs), SGSNs, and GGSNs, but also to provide support for non-3GPP access (e.g., Wi-Fi), improved policy control and charging, a wider range of QoS capabilities, advanced security/authentication mechanisms, and flexible roaming [70,71].

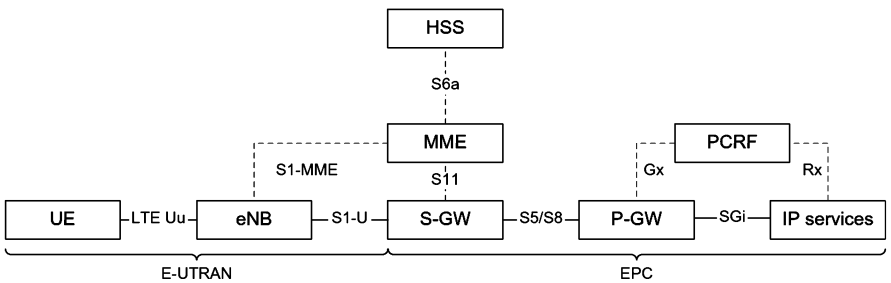


FIGURE 2.1

Generic EPS architecture [70,71].

The introduction of EPC combined with E-UTRAN in many ways represented a radical departure from previous mobile network design philosophies. It heralded the end of the circuit-switched voice-centric design era. 3GPP LTE was meant to use voice over IP (VoIP) for voice services. The EPC treats voice just as other IP-based applications, although an important one that imposes stringent performance and operational requirements [34,37,69].

3GPP LTE was required to match the wired-broadband quality of experience (QoE). This is quite different from providing mainly best-effort low-rate web browsing or short message service (SMS), two data applications for which its predecessor quasi-packet-switched mobile networks were optimized. In LTE, all mobility management functions have been moved to the core network and have become part of mobility management entity (MME). This is a consequence of the split of functions previously performed by the RNC and NB/Base Transceiver Station (BTS). The MME requires a control-plane capacity that is an order of magnitude larger than the SGSN or PDSN, and must ensure interworking with 2G/3G legacy mobile systems (see Figure 2.2).

The long-term evolution of universal mobile telecommunications system (UMTS) systems was expected to provide superior end-to-end QoS management and enforcement in order to deliver multimedia, low-latency, and real-time services, with more granular QoS classes and stricter performance targets. This was going to be achieved while ensuring scalability of users, services, and data sessions. In LTE, service control is provided via the policy and charging rules function (PCRF). This was a major change from previous systems, where service control was realized primarily through user equipment (UE) authentication by the network. The PCRF entity dynamically controls and manages all data sessions and provides appropriate interfaces toward charging and billing systems, as well as enabling new business models. The design of the new air-interface required

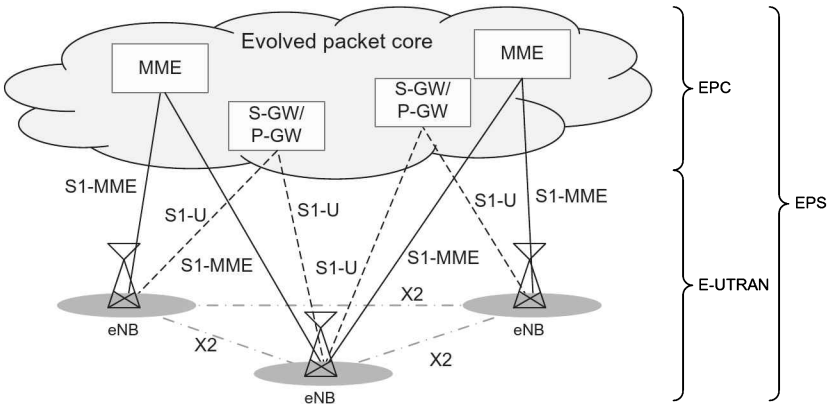


FIGURE 2.2

EPS composition [12].

significantly more capacity in the user-plane and control-plane. The existing 2G/3G mobile core elements cannot satisfactorily address LTE requirements without major upgrades to the platforms. Therefore, the legacy core network implementations must be dramatically upgraded to enable LTE.

While the core network consists of many logical nodes, the access network consists of essentially just one node, the eNB, which provides connections to the UEs. Each of the network elements is interconnected by means of logical interfaces that are standardized in order to allow multi-vendor interoperability. This gives network operators the possibility of sourcing different network elements from different vendors. In fact, network operators may choose in their physical implementations to split or merge these logical network elements depending on commercial considerations. The functional split between the EPC and E-UTRAN is shown in Figure 2.3. The EPC and E-UTRAN network elements are described in the next sections.

The EPC design has radically changed some key networking concepts which were used in the previous generations of cellular core networks and successfully addressed a number of technological challenges which had arisen from significant technological improvements in the radio access side. The LTE radio access system provides more efficient use of the spectrum with wider spectral bands reserved for the broadband systems, resulting in greater system capacity and performance.

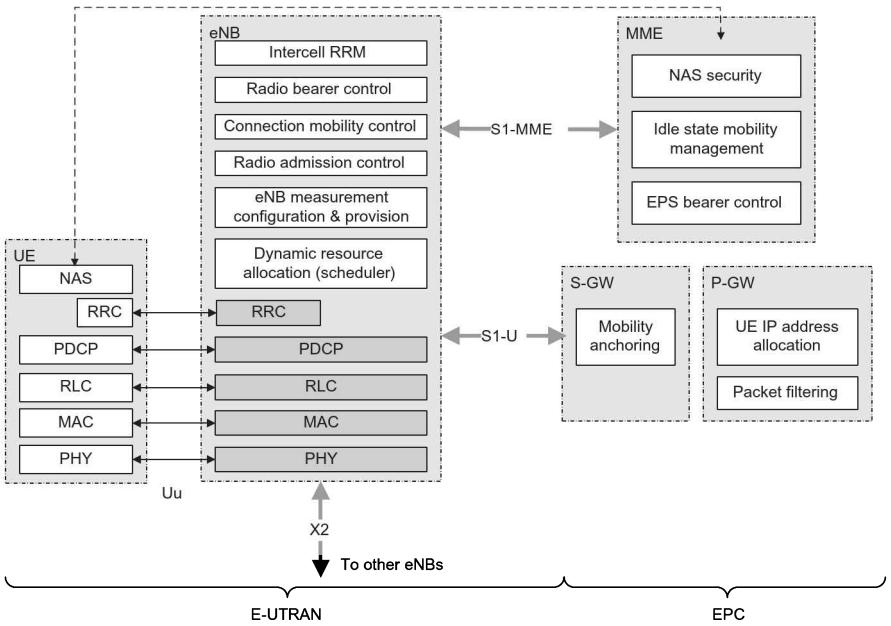


FIGURE 2.3

Functional split between E-UTRAN and EPC entities [12].

At the same time, the core network needs to change in order to provide higher throughput and lower latency; both should come as a result of simplified and improved IP-based network architecture. More specifically, the EPC addressed a number of fundamental IP-related aspects of LTE deployment including the following:

- IP routing and network addressing and real-time management
- Centralized versus distributed network architecture (for MME, serving gateway (S-GW), and packet data network gateway (P-GW) deployment)
- IPv6 (strategy, introduction, and coordination with IPv4)
- End-to-end QoS deployment and coordination with underlying transport
- Layer-2 versus layer-3 connectivity at the transport layer (eNB, S-GW, P-GW, and MME)
- End-to-end security for user-plane and control-plane
- Interconnectivity to external networks and virtual private networks
- Enabling lawful interception of data packets

There are a stringent set of requirements for reliable, scalable, high-performance network elements. Due to the dynamic nature of the user mobility in LTE, coupled with the short duration of multiple data sessions per UE and large-scale deployment targets, the EPC must further address the challenging requirements of dynamic management of mobility, policies, and signaling/data bearers, while ensuring interworking and interoperability with legacy 2G/3G systems.

The EPC is realized through the following four new network elements: (1) S-GW, (2) P-GW, (3) MME, and (4) PCRF. While S-GW, P-GW and MME were introduced in 3GPP Rel-8, PCRF had been introduced in 3GPP Rel-7. The architectures using PCRF have not been widely adopted. The PCRF interoperation with the EPC gateways and the MME is normative in 3GPP Rel-8 and essential for the operation of LTE.

The MME is a node within the EPC that performs signaling and control functions to manage the user access to network connections, assignment of network resources, and management of the mobility states to support tracking, paging, roaming, and handovers. The MME controls all control-plane functions related to subscriber and session management. The MME is the key control-node for the LTE access-network. It is responsible for idle-mode UE tracking and paging procedure. It is involved in the bearer activation/deactivation process and is also responsible for choosing the S-GW for a UE at the initial attach and at the time of intra-LTE handover involving core network node relocation. It is further responsible for user authentication. The non-access stratum (NAS) signaling terminates at the MME and it is also responsible for generation and allocation of temporary identities to UEs. It checks the authorization of the UE to camp on the service provider's public land mobile network (PLMN) and enforces UE roaming restrictions. The MME is the termination point in the network for ciphering/integrity protection for NAS signaling and handles the security key management. Lawful interception of signaling is also supported by the MME. The MME also

provides the control-plane function for mobility between LTE and 2G/3G access networks with the S3 interface terminating at the MME from the SGSN. The NAS protocols form the control-plane between the UE and the MME and support the UE mobility and the session management procedures to establish and maintain IP connectivity between the UE and a P-GW. They define the rules for a mapping between parameters during intersystem mobility with 3GPP or non-3GPP access networks. They also provide the NAS security by integrity protection and ciphering of NAS signaling messages.

The home subscriber server (HSS) is a central database that contains user- and subscription-related information. The functions of the HSS include mobility management, call and session establishment support, user authentication, and access authorization. The HSS is based on legacy home location register (HLR) and authentication center (AuC) functions in previous 3GPP releases.

The S-GW is a user-plane element whose primary function is to manage user-plane mobility and to act as a demarcation point between the radio access and core networks. The S-GW maintains data paths between eNBs and the P-GW. From a functional perspective, the S-GW is the termination point of the packet data network interface toward the E-UTRAN. When terminals move in the areas served by eNBs in the E-UTRAN, the S-GW serves as a local mobility anchor (LMA). This means that packets are routed through this point for intra E-UTRAN mobility and mobility with other 3GPP technologies, such as GSM and UMTS. The S-GW routes and forwards user data packets, while acting as the mobility anchor for the user-plane during inter-eNB handovers and as the anchor for mobility between LTE and other 3GPP radio access technologies, that is, terminating S4 interface and relaying the traffic between 2G/3G systems and P-GW. For idle mode UEs, the S-GW terminates the downlink data path and triggers paging when downlink data arrives for a UE. It manages and stores UE contexts, for example, parameters of the IP bearer service and network internal routing information. It also performs replication of the user traffic in case of lawful interception.

The P-GW is the termination point of the packet data interface toward the packet data network. As an anchor point for sessions toward the external packet data networks, the P-GW supports policy enforcement features, such as applying operator-defined rules for resource allocation and usage; packet filtering functions, such as packet inspection for application type detection and charging support. The P-GW provides connectivity from the UE to external packet data networks by being the point of exit and entry of traffic for the UE. A UE may have simultaneous connections with more than one P-GW for accessing multiple packet data networks (PDNs). Another key role of the P-GW is to act as the anchor for mobility between 3GPP and non-3GPP technologies, such as CDMA2000.

In LTE, the data-plane traffic is carried via virtual connections known as service data flows (SDFs). The SDFs are carried over bearers that are virtual containers with unique QoS characteristics. Figure 2.4 illustrates an example scenario

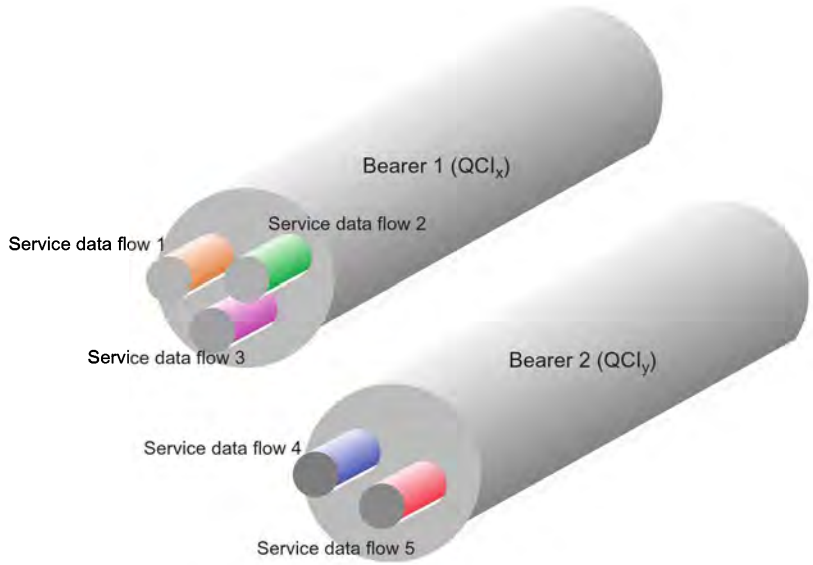


FIGURE 2.4

Relationship between SDFs and bearers [70,71].

where one or more SDFs are aggregated and assigned to one bearer. One bearer, a data path between a UE and a PDN, has three segments as follows: radio bearer between UE and eNB; data bearer between eNB and S-GW (S1 bearer); and data bearer between S-GW and P-GW (S5 bearer). Figure 2.5 illustrates three segments that constitute an end-to-end bearer. The primary role of a P-GW is QoS enforcement for each of these SDFs, while S-GW dynamically manages the bearers. There are two types of EPS bearers: the *default EPS bearer* and the *dedicated EPS bearer*. The default EPS bearer is established during the attach procedure and allocates an IP address to the UE that does not have a specific QoS (only a nominal QoS). A dedicated EPS bearer is typically established during call setup and after transition from idle mode to the connected mode. It does not allocate any additional IP address to the UE and is linked to a specified default EPS bearer but has a specific QoS class. A QoS class identifier (QCI) is a scalar that is used as a reference to access node-specific parameters that control bearer level packet forwarding (e.g., scheduling weights, admission thresholds, queue management thresholds, and link layer protocol configuration), and are preconfigured by the operator owning the access node (eNB).

As shown in Figure 2.5, the EPS bearer transports packets by an S1 bearer between the S-GW and the eNB, and by a radio bearer between a UE and an eNB. There is a one-to-one correspondence between a radio bearer identifier and an S1 bearer. The IP packets mapped to the same EPS bearer receive the same bearer-level packet forwarding treatment, e.g., scheduling policy, queue management policy, rate shaping policy, and radio link control (RLC) configuration. In order to

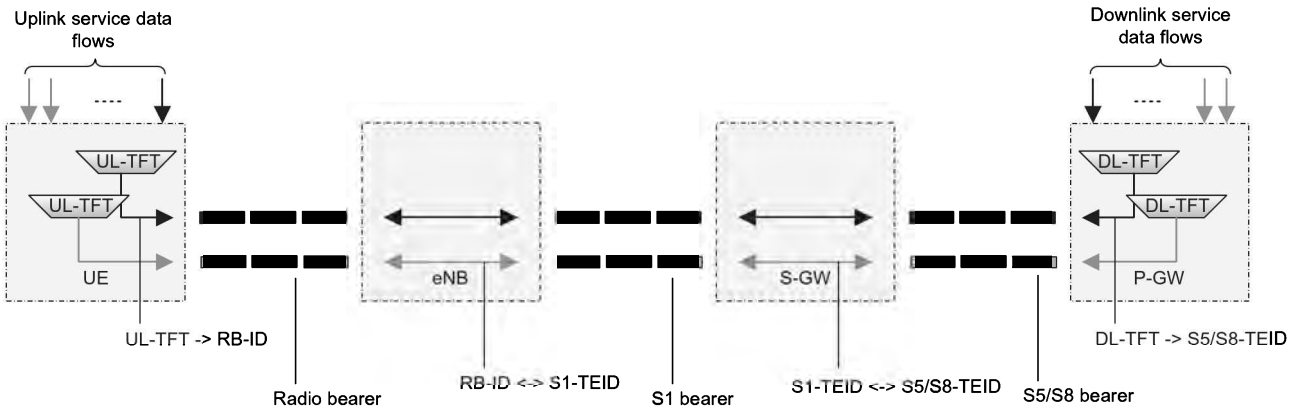


FIGURE 2.5

Bearers across different EPS interfaces [70,71].

provide different bearer-level QoS, a separate EPS bearer should be established. The IP packets must then be filtered and mapped to the appropriate EPS bearers. Packet filtering of different bearers is based on traffic flow templates (TFTs). The TFTs use IP header information, such as source and destination IP addresses and transport control protocol (TCP) port numbers to filter packets, such as VoIP from web-browsing traffic, so that each flow can be mapped to the appropriate EPS bearers with the appropriate QoS settings. An uplink TFT (UL-TFT) associated with each bearer in the UE filters IP packets to EPS bearers in the uplink direction. A downlink TFT (DL-TFT) in the P-GW is a similar set of downlink packet filters. When a UE attaches to the network, the UE is assigned an IP address by the P-GW and a default bearer is established, and it remains in effect during the lifetime of the PDN connection in order to provide the UE with always-on connectivity to that PDN. The initial bearer-level QoS parameter values of the default bearer are assigned by the MME based on subscription data retrieved from the HSS. The policy and charging enforcement function (PCEF) may change these values in interaction with the PCRF or according to local configuration. One or more dedicated bearers can be established at any time during or after completion of the attach procedure. A dedicated bearer can be either a guaranteed bit rate (GBR) or a non-GBR bearer. The default bearer is always a non-GBR bearer. The distinction between default and dedicated bearers should be transparent to the access network. Each bearer has an associated QoS, and if more than one bearer is established for a given UE, then each bearer must also be associated with appropriate TFTs. The dedicated bearers may be established by the network based on a trigger from the IP multimedia subsystem (IMS) domain or a request by the UE. The dedicated bearers for a UE may be provided by one or more P-GWs.

In the E-UTRAN side, the evolved nodeB (eNB) provides the E-UTRA user-plane and control-plane protocol termination toward the UE. The eNBs can optionally be interconnected via X2 logical interfaces. The X2 interface may not be necessarily in the form of a point-to-point physical connection between eNBs. The eNBs are connected by means of the S1 interface to the EPC. More specifically, it is connected to the MME by means of the S1-MME reference point, and to the S-GW by means of the S1-U interface. The S1 interface supports a many-to-many relationship between MMEs/S-GWs and eNBs.

2.1.2 Evolution of the network architecture

The global system for mobile communications (GSM) network architecture was based on a circuit-switched design where a direct connection was established between the call-making and call-receiving entities throughout the telecommunication network, that is, radio access and core networks of the mobile operator and the fixed network. The GSM was a voice-centric system providing circuit-switched telephony, SMS, and some very low-rate data services. The general packet radio service (GPRS) enhancement added packet-switched services to the circuit-switched functions, where data was transported in the form of packets without

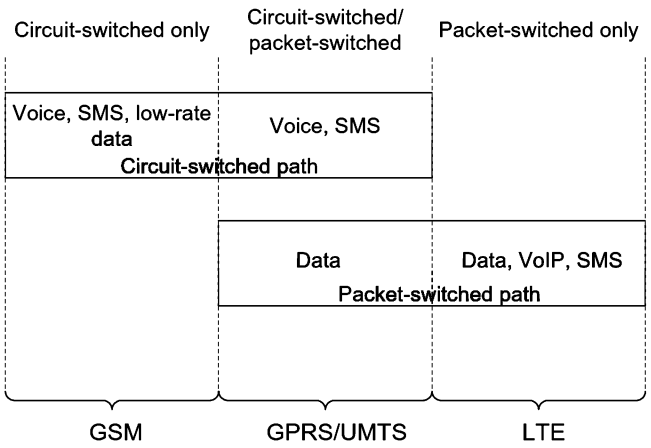


FIGURE 2.6

Evolution of cellular networks and services.

establishing dedicated circuits. This enhancement offered more flexibility and efficiency. In GPRS, voice and SMS are transported via circuit-switched paths; thus, the core network is composed of circuit-switched and packet-switched domains. In UMTS, the dual-domain concept was maintained on the core network side and only a few network elements evolved. In the design of the 3G system long-term evolution, the 3GPP decided to use the IP as the key protocol to transport all services; therefore, the EPC vision did not include a circuit-switched domain and it was an evolution of the packet-switched architecture used in GPRS/UMTS. This paradigm shift in the network design not only had significant impact on the architecture itself but also on the way that the services were provided. Traditional use of circuits to carry voice and short messages needed to be replaced by IP-based solutions in the long term. The evolution of 3GPP network architecture and services are depicted in Figures 2.6–2.8 from different perspectives (services, general architecture, and network elements). As shown in the figures, while the services have migrated toward pure IP-based services, the network architecture has been simplified and redundant entities have been eliminated or integrated into other network elements.

Different options are possible to integrate MME, S-GW, and P-GW logical entities. Each option will result in different handling of control-plane and user-plane packet routing. The choice may depend on operator strategy and/or EPC vendor equipment availability and constraints. Integrating more logical entities together in one physical entity will reduce latency/delay, cost (fewer nodes and interfaces) and signaling load, but will increase the functional complexity per entity/component and will limit the benefit and potential savings achieved by a split architecture, i.e., redundancy, flexibility, and scalability. A possible deployment option may start with some level of integration, to save initial costs, and migrate to more flexible and scalable architecture when the number of users and user traffic grows.

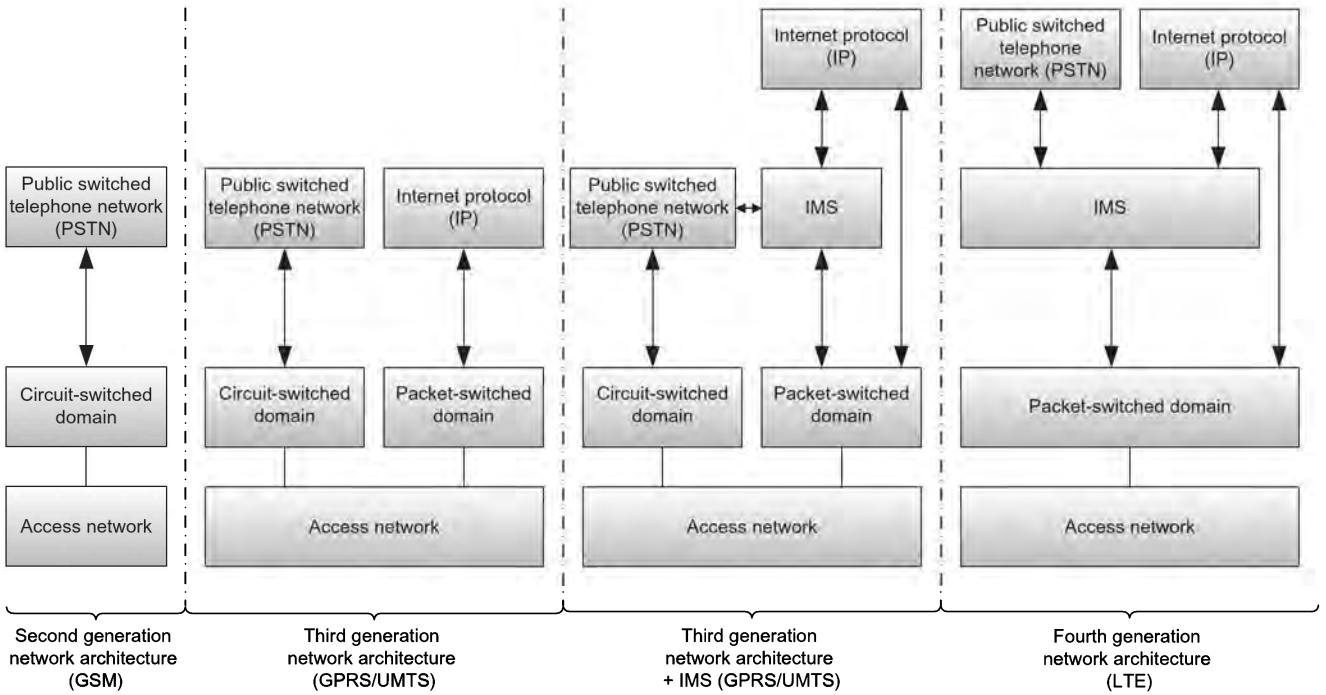


FIGURE 2.7

Evolution of network architecture through cellular generations.

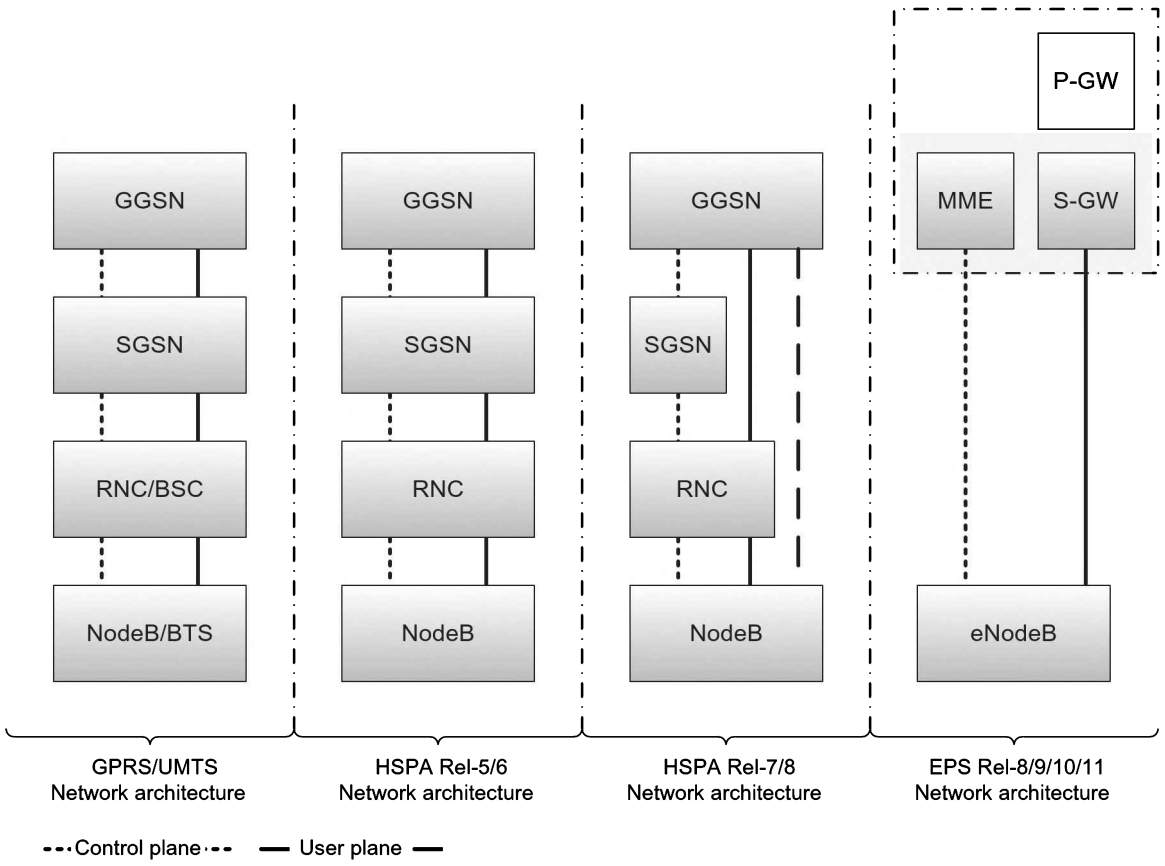


FIGURE 2.8

Evolution of 3GPP network architecture through different releases.

2.2 EPC architecture

2.2.1 Reference model and network interfaces

As mentioned earlier, the EPC is an all-IP mobile core network for the LTE radio access, providing a converged framework for IP-based real-time/non-real-time applications and services. The EPC provides the entire core-network functionality that, in previous mobile network generations, was realized through two separate domains: circuit-switched for voice and packet-switched for data. As shown in Figure 2.7, in LTE, these two distinct mobile core network domains used for separate processing and switching of voice and data are unified in a single IP-based domain. The EPC implementation and deployment is essential for end-to-end IP-based service delivery which complements the LTE radio access network. The EPC promotes the introduction of new services and the enablement of new applications [70,71].

Figure 2.9 shows the EPS reference model, which is a logical representation of the network architecture. The network reference model identifies the functional entities in the architecture and the reference points between the functional elements over which interoperability is achieved. The overall architecture has two distinct components: the access network and the core network. The access network is the E-UTRAN and the core network is known as the EPC. Voice services which were traditionally circuit-switched are handled using IMS infrastructure. Network complexity and latency are reduced as there are fewer nodes in both signaling and data paths. The EPC is further designed to support non-3GPP access for mobile IP. To improve system robustness, security, integrity protection, and ciphering have been added and represented by NAS, which is an additional layer of abstraction to protect important information, such as key and security interworking between 3GPP and non-3GPP networks. Apart from the network entities handling user-plane traffic, the EPC further includes network control entities for storing user subscription information represented by HSS, determining the identity and privileges of a user and tracking activities via authorization, authentication and accounting (AAA) server, and enforcing charging and QoS policies through PCRF. Note that E-UTRAN and EPC together constitute the EPS. The radio access and core networks perform several functions including network access control functions; packet

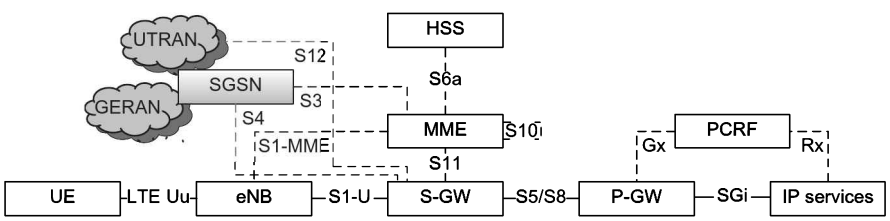


FIGURE 2.9

EPS reference model [6].

routing and transfer functions; mobility management functions; security functions; RRM functions; and network management functions.

Some general principles that were taken into consideration in the design of E-UTRAN architecture as well as the E-UTRAN interfaces are summarized as follows:

- Signaling and data transport networks are logically separated.
- E-UTRAN and EPC functions are fully separated from transport functions. The addressing schemes used in E-UTRAN and EPC are not associated with the addressing schemes of transport functions. The fact that some E-UTRAN or EPC functions may reside in the same physical equipment does not make the transport functions part of the E-UTRAN or the EPC.
- The mobility for RRC connection is fully controlled by the E-UTRAN.
- The interfaces are based on the logical model of the network entities.
- One physical network element can implement multiple logical nodes.

3GPP network architecture defines a reference point as a logical link that connects two groups of functions that reside in different functional entities of the E-UTRAN or EPC. Figure 2.9 illustrates the EPS network reference model and the most commonly used reference points specified by 3GPP as of Rel-11. These reference points are defined in Table 2.1.

2.2.2 Public land mobile network

A PLMN provides land mobile telecommunication services to the public which is established and operated by an administration or private operating agency. A PLMN may be referred to as an extension of networks, e.g., integrated services data digital network (ISDN), corporate, and public PDNs. The PLMN is a collection of mobile switching centers (MSCs) in circuit-switched domain and SGSNs in GPRS core network or SGSNs/MMEs in EPC, in packet-switched domain under a common numbering and routing plan. The MSCs are the functional interfaces between the fixed networks and a PLMN for call setup in circuit-switched domain. The GGSN and the SGSN are the functional interfaces between the fixed telephone networks and a PLMN for packet transmission in a GPRS packet-switched domain. In the case of an EPC packet-switched domain, the P-GW, S-GW, SGSN, and the MME are the functional interfaces between the fixed networks and a PLMN for packet transmission. From a functional point of view, the PLMNs may be considered as independent telecommunication entities even though different PLMNs may be interconnected through an ISDN or public switched telephone network (PSTN) and PDNs for the purpose of forwarding calls or network information. A similar type of interconnection may exist for interaction between the MSCs, SGSNs, or MMEs of a PLMN [1].

The PLMN infrastructure typically consists of a core network and an access network. The core network is logically divided into a circuit-switched domain, a packet-switched domain, and an IMS domain. The access network may include a

Table 2.1 EPS Reference Points [6,33]

Reference Point	Termination Points	Description
S1-MME	eNB and MME	It provides control-plane path between eNB and MME.
S1-U	eNB and S-GW	It enables per-bearer user-plane tunneling and inter-eNB path switching during handover. The S1-U is based on GTP-U protocol.[a]
S3	MME and SGSN	It enables user and bearer information exchange for inter-3GPP access network mobility in idle and/or active mode. This reference point can be used intra-PLMN or inter-PLMN in the case of inter-PLMN handover. The S3 is based on GTP protocol.
S4	S-GW and SGSN	It provides control and mobility support between GPRS core and the 3GPP anchor function of S-GW and is based on GTP protocol. If direct tunneling is not established, it provides the user-plane tunneling capability.
S5	S-GW and P-GW	It provides user-plane tunneling and tunnel management between S-GW and P-GW. It is used for S-GW relocation due to UE mobility and if the S-GW needs to connect to a non-collocated P-GW for PDN connectivity and is based on GTP protocol.
S6a	MME and HSS	It enables transfer of subscription and authentication data for authenticating/authorizing user access to the network services.
Gx	PCRF and P-GW	It provides transfer of QoS policy and charging rules from PCRF to PCEF in the P-GW.
S8	S-GW and P-GW	Inter-PLMN reference point providing user-plane and control-plane between the S-GW in the visited PLMN and the P-GW in the home PLMN. S8 is the inter-PLMN variant of S5 and is based on GTP protocol.
S9	Visited PCRF and home PCRF	It provides transfer of QoS PCC information between the home PCRF and the visited PCRF in order to support local breakout function.
S10	MME and MME	It is used for MME relocation and MME to MME information transfer. This reference point can be used in the form of intra-PLMN or inter-PLMN in the case of inter-PLMN handover.
S11	MME and S-GW	It provides the reference point between MME and S-GW.
S12	UTRAN and S-GW	It provides user-plane tunneling when direct tunnel is established.
S13	MME and EIR	It enables UE identity check procedure between MME and equipment identity register (EIR).[b]

(*Continued*)

Table 2.1 (Continued)

Reference Point	Termination Points	Description
SGi	P-GW and PDN	It provides the interface between the P-GW and the packet data network. Packet data network may be an operator external public or private packet data network or an intra-operator packet data network as a provision for IMS services.
Rx	PCRF and PDN	This reference point resides between the application function (AF) and the PCRF.

[a]GPRS tunneling protocol (GTP) is a group of IP-based communications protocols used to carry GPRS within GSM, UMTS, and LTE networks [35]. In 3GPP architecture, GTP and PMIPv6-based interfaces are specified on various reference points. GTP can be decomposed into separate protocols, GTP-C, GTP-U, and GTP'. GTP-C is used within the GPRS core network for signaling between gateway GPRS support node (GGSN) and serving GPRS support node (SGSN). This allows the SGSN to activate a session on a user's behalf (PDP context activation), to deactivate the same session, to adjust QoS parameters, or to update a session for a subscriber joined from another SGSN. GTP-U is used for carrying user data within the GPRS core network and between the radio access network and the core network. The user data transported can be packets in any of IPv4, IPv6, or PPP formats. The primary bearer plane in LTE across the access and the core networks is the evolved GPRS tunneling protocol (eGTP). The user-plane variant of this protocol eGTP-U provides tunneling support for user data between the eNB and the EPC elements. The eGTP-C (GTPv2-C) protocol is responsible for creating, maintaining, and deleting tunnels on multiple Sx interfaces. It is used for control-plane path management, tunnel management, and mobility management. It also controls forwarding relocation messages, serving radio network subsystem (SRNS) context and creating forward tunnels during inter-LTE handovers [9].

[b]The EIR is often integrated to the HLR. The EIR maintains a list of mobile phones (identified by their IMEI) which are banned from the network or need to be monitored. This is designed to allow tracking of stolen mobile phones. The EIR is a database that contains information about the identity of the mobile equipment that prevents calls from stolen, unauthorized or defective mobile stations [51].

single or multiple radio access networks, such as GERAN, UTRAN, or E-UTRAN. The core network, on the other hand, consists of the circuit-switched and packet-switched domains, where the latter includes GPRS core and EPC. These two domains differ in the way that they support user traffic and may include a number of common entities. A PLMN can implement only one domain or both domains. The circuit-switched domain refers to the set of core network entities providing circuit-switched type of connection for user applications and the related signaling. A circuit-switched connection is a connection for which dedicated network resources are allocated at the time of connection establishment and are released at the end of connection.

The packet-switched domain refers to the set of core network entities that provide a packet-switched type of connection for user applications and the related signaling. A packet-switched connection transports user information in the form of bit packets, where each packet can be routed independently from the others.

The network entities specific to the packet-switched domain are the GPRS support entities, i.e., SGSN and GGSN, and EPS-specific entities comprising P-GW, S-GW, and MME. The IMS domain encompasses those core network elements which enable provisioning and delivery of IP-based multimedia services, such as voice, audio, video, text, or chat over the packet-switched domain.

2.2.3 Packet data network gateway

The P-GW is the entity which terminates the SGi interface toward the packet data network. If a UE has multiple PDN connections, then the UE may be assigned to more than one P-GW. The P-GW provides PDN connectivity to GERAN/UTRAN capable UEs as well as E-UTRAN capable UEs using E-UTRAN, GERAN, or UTRAN access networks. The P-GW provides PDN connectivity to E-UTRAN capable UEs using E-UTRAN only over the S5/S8 interface. It may also provide PDN connectivity to UEs using non-3GPP access networks. The P-GW functions include per-user packet filtering; lawful interception; IP address allocation; transport-level packet marking in the uplink and downlink by setting the differentiated services code point based on the QCI of the associated EPS bearer; uplink/downlink service level charging, gating control, rate enforcement; uplink/downlink rate enforcement based on per access point name-aggregate maximum bit rate (APN-AMBR)[1] parameter; downlink rate enforcement based on the accumulated maximum bit rates of a set of SDFs with the same GBR QCI through rate policing/shaping; and dynamic host configuration protocol (DHCP)v4 (server and client) and DHCPv6 (client and server) functions.[2] In addition, the P-GW includes uplink/downlink bearer binding (bearer binding is the generic procedure for associating a bearer in the access network to an SDF) and uplink bearer binding verification functions for the GTP-based network interfaces. The P-GW functions also include user-plane anchor for mobility between 3GPP access and non-3GPP access.

[1]The APN-AMBR is a subscription parameter stored per APN in the HSS. It limits the aggregate bit rate that can be provided across all non-GBR bearers and across all PDN connections of the same APN (e.g., excess traffic may be discarded by a rate shaping function). Each of those non-GBR bearers could potentially utilize the entire APN-AMBR, for example, when the other non-GBR bearers do not carry any traffic. The GBR bearers are outside the scope of APN-AMBR. The P-GW enforces the APN-AMBR in downlink. Enforcement of APN-AMBR in uplink is done in the UE and additionally in the P-GW [1].

[2]DHCP is a network protocol that is used to configure devices that are connected to a network so they can communicate on that network using the Internet protocol. It involves clients and a server operating in a client-server model. The clients request configuration settings using the DHCP protocol, such as an IP address, a default route, and one or more domain name system (DNS) server addresses. Once the client implements these settings, the host is able to communicate on the Internet. The DHCP server maintains a database of available IP addresses and configuration information. When the server receives a request from a client, the DHCP server determines the network to which the DHCP client is connected, and then allocates an IP address or prefix that is appropriate for the client, and sends configuration information appropriate for that client. DHCP servers typically grant IP addresses to clients only for a limited interval. DHCP is used for IPv4 and IPv6 [27,74].

2.2.4 Serving gateway

The S-GW is the entity which terminates the interface toward E-UTRAN access network. For each UE associated with the EPS, at any given point of time, there is a single S-GW assigned. The functions of the S-GW include a local mobility anchor point for inter-eNB handover; mobility anchor point for inter-3GPP mobility; EPS connection management (ECM)-IDLE mode downlink packet buffering and initiation of network-triggered service request procedure; lawful interception; packet routing and forwarding; transport-level packet marking in the uplink and downlink; and accounting for interoperator charging; as well as local non-3GPP anchor for the case of roaming when the non-3GPP IP access network is connected to the visited PLMN; event reporting (change of RAT) to the PCRF; uplink and downlink bearer binding toward 3GPP access; uplink bearer binding verification with packet dropping of misbehaving uplink traffic; mobile access gateway (MAG) functions, if proxy mobile IP (PMIP)[3] based S5 or S8 is used; and support of necessary functions for enabling GTP/PMIP chaining functions [6,7].

2.2.5 Mobility management entity

The MME is the main control node that processes the NAS signaling between the UE and the core network which is responsible for idle-mode UE tracking and paging procedure as well as bearer establishment and release [10]. The main functions supported by the MME can be classified into two major categories: (1) functions related to bearer management including the establishment, maintenance, and release of the bearers which are handled by the session management function in the NAS protocol; and (2) functions related to connection management including the establishment of the connection and security association (SA) between the network and UE which is handled by the connection or mobility management function in the NAS protocol [11]. The MME is a control-plane entity within EPS which performs mobility management functions including NAS signaling and security; inter-node signaling for mobility between 3GPP access networks; tracking area list management; P-GW and S-GW selection; SGSN selection for handovers to 2G/3G access networks; roaming; and authentication. The MME's bearer management functions include dedicated bearer establishment. In order to enable 3GPP2 access (in a multi-RAT network), the MME should support high rate packet data (HRPD) access node selection and maintenance for handovers to HRPD and transparent transfer of HRPD signaling messages and transfer of status information between E-UTRAN and HRPD access networks [1].

[3]Network-based mobility management enables IP mobility for a host without requiring its participation in any mobility-related signaling. The network is responsible for managing IP mobility on behalf of the host. The mobility entities in the network are responsible for tracking the movements of the host and initiating the required mobility signaling on its behalf. IETF RFC 5213 specifies a network-based mobility management protocol that is referred to as proxy mobile IPv6 [26].

The MME creates a UE context when a UE is powered on and attaches to the network. It assigns a unique short temporary identity to the UE that identifies the UE context in the MME. This UE context holds user subscription information downloaded from the HSS. The local storage of subscription data in the MME allows faster execution of procedures, such as bearer establishment since it relaxes the need to frequently exchange messages with the HSS. In addition, the UE context also holds dynamic information, such as the list of bearers that are established and the terminal capabilities [70,71].

To reduce the overhead in the E-UTRAN and processing complexity in the UE, all UE-related information in the access network including the radio bearers can be released during long periods of user inactivity, which is known as ECM-IDLE state. The MME retains the UE context and the information about the established bearers during these idle periods. To allow the network to contact an idle-mode UE, the UE provides the network with location update whenever it moves out of its current tracking area using a tracking area update procedure. The MME is responsible for tracking user location while the UE is in the ECM-IDLE state. Upon availability of downlink data for an ECM-IDLE UE, the MME sends a paging message to all eNBs in the tracking area where the UE was recently located and the eNBs page the UE over the radio interface. Upon receipt of the paging message, the UE performs a service request procedure, which results in transitioning to the ECM-CONNECTED state. The UE-related information is thereby created in the E-UTRAN and the radio bearers are reestablished. The MME is responsible for the reestablishment of the radio bearers and updating the UE context in the eNB (see Chapters 4 and 5 for more details).

Security functions are the responsibility of the MME for both signaling and user data. When a UE attaches to the network, a mutual authentication of the UE and the network is performed between the UE and the MME/HSS. This authentication procedure establishes the security keys that are used for encryption of the bearers.

2.2.6 Policy and charging rules function

The PCRF acts as a policy decision point for policy and charging control (PCC) of SDFs/applications and IP bearer resources. It was introduced as an improvement in 3GPP Rel-7 network architecture in terms of policy and charging to allow optimization of interactions between the policy and rules functions. This improvement involved a new network node which integrates policy decision function (PDF) and charging rules function (CRF) in the previous releases. 3GPP Rel-8 further enhanced PCRF functionality by expanding the scope of PCC framework to facilitate non-3GPP access to the network (e.g., Wi-Fi or fixed broadband access). In the 3GPP generic PCC model, policy and charging enforcement function (PCEF), which resides in the P-GW, is the functional entity that supports SDF detection, policy enforcement, and flow-based charging. The PCRF provides the QoS authorization (QCI and bit rates) that decides how a certain data flow

will be treated in PCEF and ensures consistency with the user's subscription profile. The application function (AF) represents the network element that supports applications that require dynamic policy and/or charging control.

2.2.7 Home subscriber server

The home subscriber server is the subscription data repository for all permanent user data. The HSS stores the master copy of the subscriber profile, which contains information about the services that are applicable to the user. It also records the location of the user at the level of visited network control node, such as MME. It is the entity containing the subscription-related information to support the network entities that handle calls/sessions. A home network may contain one or several HSSs depending on the number of mobile subscribers, the capacity of the equipment, and the organization of the network. As an example, the HSS provides support to the call control servers in order to complete the routing/roaming procedures by solving authentication, authorization, naming/addressing resolution, location dependencies, etc. The HSS is responsible for holding information related to user identification, numbering and addressing information; user security information; network access control information for authentication and authorization; and user location information at the inter-system level. The HSS supports the user registration and stores inter-system location information as well as the user profile information. The HSS also generates user security information for mutual authentication, communication integrity check, and ciphering. Based on this information, the HSS is also responsible for supporting the call control and session management entities of the different domains and subsystems [1].

2.2.8 Authorization, authentication, and accounting

Trusted and untrusted non-3GPP access networks are IP-based access networks which may use access technologies whose specifications are out of the scope of 3GPP. Whether a non-3GPP access network is trusted or untrusted is not a characteristic of the access network. In non-roaming scenario, it is the home PLMN's operator decision if a non-3GPP IP-based access network is treated as trusted or untrusted. In a roaming scenario, the HSS/3GPP AAA server in home PLMN determines whether a non-3GPP IP access network is viewed as trusted or untrusted. The HSS/3GPP AAA server may consider and utilize the visited PLMN's policy and capability returned from the 3GPP AAA proxy or roaming agreement.

The 3GPP AAA server is located in the home PLMN and provides user support for non-3GPP access with authentication, authorization, and location management in order to allow access to the EPS. It also contains necessary user-related information in order to grant access to a non-3GPP network, and further coordinates the information needed to support mobility between 3GPP and non-3GPP access networks, such as coordination of P-GW information. It interacts with HSS

to maintain consistent information for users supporting mobility and service continuity between 3GPP and non-3GPP networks. The 3GPP AAA proxy provides support for roaming non-3GPP access users in the visited PLMN with the necessary authentication, authorization, and location management services in order to get access to the EPS. It may also provide roaming-related information for support of chaining scenarios. If an S-GW is needed for non-3GPP access in the visited network, the 3GPP AAA proxy selects an S-GW for the UE during initial attach or handover process [2].

2.2.9 EPC identifiers

Each entity and bearer in E-UTRAN and EPC is identified with a unique identifier. The identifiers are either permanently provisioned, such as the IMSI and IMEI identities that are assigned by an operator to the UE or they are assigned during the lifetime of UE operation in an operator's network. Figure 2.10 illustrates the identities that are either provisioned or dynamically/semi-statically assigned to the bearers and the entities in EPS. These identities are described in Table 2.2 and the next chapters of this book and further specified by 3GPP in 3GPP TS 23.003 specification [2]. The direction of the arrows in the figure indicates the entity which assigns the identifier and the entity to which the identifier is assigned. The bearers and various bearer identifiers depending on the network interface and the flow direction are also shown in the figure.

Note that the protocols over LTE-Uu and S1 interfaces are divided into two categories: user-plane protocols, which are the protocols implementing the actual E-RAB service carrying user data through the access stratum (AS); and control-plane protocols that are the protocols for controlling the E-RABs and the connection between the UE and the network from different aspects including requesting the service, controlling different transmission resources, handover, etc. Furthermore, a mechanism for transparent transfer of NAS messages is included.

2.3 Roaming architecture and interworking with other networks

In cellular communications, roaming is a general term referring to the extension of service continuity and availability in a location that is different from the home location where the service and the subscriber were registered. Roaming functionality ensures that the wireless device is connected to the network without connection interruption while crossing the PLMN boundaries. The roaming is defined as the ability for a cellular user to automatically make or receive voice calls; send or receive data; or access other services including home data services, when moving outside the geographical coverage area of the home network using a visited network. The roaming feature is technically supported by mobility management,

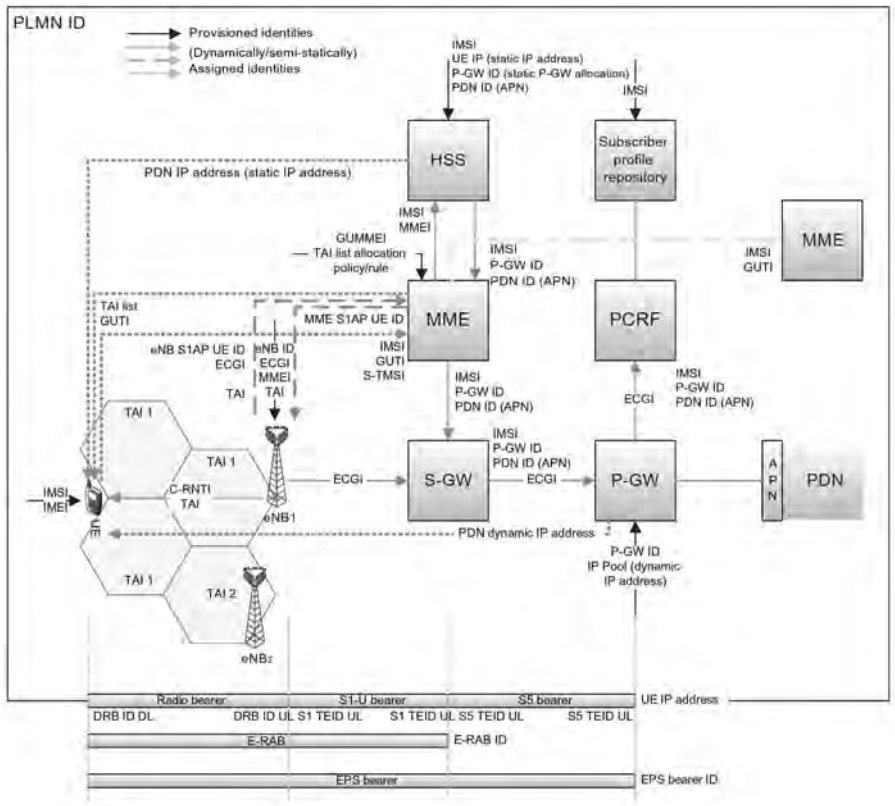


FIGURE 2.10

EPS entities (bearers) and corresponding provisioned/assigned identities [40].

authentication, authorization, and billing functions in the network. While a roaming user is connected to the E-UTRAN, MME, and S-GW of the visited LTE network, the EPS allows the P-GW of either the visited or the home network to be used, as shown in Figure 2.11. Using the home network's P-GW allows the user to access the home operator's services while in a visited network. A P-GW in the visited network allows a local breakout to the Internet in the visited network [70,71].

The EPS also supports interworking with and handover to/from other networks using 3GPP/non-3GPP radio access technologies, such as GSM, UMTS, and CDMA2000. The architecture for interworking with 2G/3G GPRS/UMTS networks is shown in Figure 2.12. The S-GW acts as the mobility anchor for interworking with other 3GPP technologies, such as GSM and UMTS, while the P-GW serves as an anchor allowing seamless mobility to non-3GPP networks, such as CDMA2000 or mobile WiMAX. The P-GW may also support a PMIP-based interface.

able 2.2 Description of EPC and E-UTRAN Identities [40,41]

bbreviation	Identifier	Description	Value/Structure	Permanent/ Temporary
MSI	International mobile subscriber identity	Unique identification of mobile subscriber in the network. The IMSI is composed of three parts: (1) mobile country code (MCC) consisting of three digits. The MCC uniquely identifies the country of residence of the mobile subscriber; (2) mobile network code (MNC) consisting of two or three digits for GSM/UMTS applications. The MNC identifies the home PLMN of the mobile subscriber. The length of the MNC (two or three digits) depends on the value of the MCC. A mixture of two and three digit MNC codes within a single MCC area is not recommended and is outside the scope of this specification; and (3) mobile subscriber identification number (MSIN) identifying the mobile subscriber within a PLMN.	IMSI (not more than 15 digits) = PLMN ID + MSIN = MCC + MNC + MSIN	Permanent
LMN ID	Public land mobile network identifier	Unique identification of PLMN	PLMN ID (not more than 6 digits) = MCC + MNC	Permanent
MCC	Mobile country code	Assigned by ITU	3 digits	Permanent
MNC	Mobile network code	Assigned by national authority	2 or 3 digits	Permanent
MSIN	Mobile subscriber identification number	Assigned by operator	9 or 10 digits	Permanent

GUTI	Globally unique temporary UE identity	Identifies a UE (between the UE and the MME) on behalf of IMSI for security reasons. The purpose of the GUTI is to provide an unambiguous identification of the UE that does not reveal the UE or the user's permanent identity in the EPS. It also allows the identification of the MME and network. It can be used by the network and the UE to establish the UE's identity during signaling in the EPS.	GUTI (not more than 80 bits) = GUMMEI + M-TMSI	Temporary
TIN	Temporary identity used in next update	GUTI is stored in TIN parameter of UE's context. TIN indicates which temporary ID will be used in the next update	TIN = GUTI	Temporary
S-TMSI	SAE temporary mobile subscriber identity	To locally identify a UE in short-form within an MME group (unique within an MME pool). The S-TMSI is the shortened form of the GUTI to enable more efficient radio signaling procedures (e.g., paging and service request). For paging purposes, the mobile is paged with the S-TMSI.	S-TMSI (40 bits) = MMEC + M-TMSI	Temporary
M-TMSI	MME mobile subscriber identity	Unique within an MME Since the TMSI has only local significance (i.e., within a VLR and the area controlled by a VLR, or within an SGSN and the area controlled by an SGSN, or within an MME and the area controlled by an MME), the structure and coding of it can be chosen by agreement between operator and manufacturer in order to meet local needs.	32 bits	Temporary
GUMMEI	Globally unique MME identity	To identify an MME uniquely in a global sense.	GUMMEI (not more than 48 bits) = PLMN ID + MMEI	Permanent
MMEI	MME identifier	To identify an MME uniquely within a PLMN	MMEI (24 bits) = MMEGI + MMEC	Permanent

(*Continued*)

able 2.2 (Continued)

bbreviation	Identifier	Description	Value/Structure	Permanent/ Temporary
MMEGI	MME group identifier	Unique identification of an MME group within a PLMN	16 bits	Permanent
MMEC	MME code	To identify an MME uniquely within an MME group	8 bits	Permanent
MEI	International mobile equipment identity	To uniquely identify a mobile equipment	IMEI (15 digits) = type approval code (6 digits) + serial number (6 digits) + final assembly code (2 digits) + check digit (1 digit)	Permanent
MEI/SV	IMEI/software version	Identifies the mobile equipment serial number, hardware model number, and software version number	IMEI/SV (16 digits) = type approval code (6 digits) + serial number (6 digits) + software version (2 digits) + final assembly code (2 digits)	Permanent
CGI	E-UTRAN cell global identifier	To identify a cell in general (globally unique) EPC can know UE location based on ECGI	ECGI (not more than 52 bits) = PLMN ID + ECI	Permanent
CI	E-UTRAN cell identifier	To identify a cell within a PLMN	ECI (28 bits) = eNB ID + cell ID	Permanent
lobal eNB)	Global eNB identifier	To globally identify an eNB	PLMN ID + eNB ID	Permanent
NB ID	eNB identifier	To identify an eNB within a PLMN	20 bits	Permanent
-GW ID	PDN GW identity	To identify a specific P-GW HSS assigns P-GW for PDN (IP network) connection of each UE	IP address (4 bytes) or fully qualified domain name (variable length)	Permanent
AI	Tracking area identity	This is the identity used to uniquely identify the tracking areas. The tracking area identity is constructed from the MCC (mobile country code), MNC (mobile network code) and TAC (tracking area code).	TAI (not more than 32 bits) = PLMN ID + TAC P-GW	Permanent

AC	Tracking area code	To indicate to which tracking area the eNB belongs (per cell) which is unique within a PLMN	16 bits	Permanent
AI list	Tracking area identity list	UE can move into the cells included in TAI list without location update (TA update) which is globally unique	Variable length	Permanent
DN ID	Packet data network identity	To identify a PDN (IP network) with which mobile data user communicates PDN identity (APN) is used to determine the P-GW and point of interconnection with a PDN With APN as query parameter to the DNS procedures, the MME will receive a list of candidate P-GWs, and then a P-GW is selected by MME	PDN identity = APN = APN-NI + APN-OI (variable length) APN-NI: APN network identifier APN-OI: APN operator identifier	Permanent
SG ID	Closed subscriber group ID	A closed subscriber group (CSG) consists of a single cell or a collection of cells within the E-UTRAN and UTRAN that are open to only a certain group of subscribers. A CSG ID is a unique identifier within the scope of PLMN which identifies a CSG in the PLMN associated with a CSG cell or group of CSG cells.	27 bits	Permanent
PS bearer ID	Evolved packet system bearer identifier	To identify an EPS bearer (default or dedicated) per UE. An EPS bearer identity uniquely identifies an EPS bearer for one UE accessing via E-UTRAN. The EPS bearer identity is allocated by the MME. There is a one-to-one mapping between EPS radio bearer and EPS bearer, and the mapping between EPS radio bearer identity and EPS bearer identity is made by E-UTRAN. The E-RAB ID value, used over S1 and X2 interfaces to identify an E-RAB, is the same as the EPS bearer ID value used to identify the associated EPS bearer.	4 bits	Temporary

(*Continued*)

able 2.2 (Continued)

bbreviation	Identifier	Description	Value/Structure	Permanent/ Temporary
-RAB ID	E-UTRAN radio access bearer identifier	To identify an E-RAB per UE	4 bits	Temporary
RB ID	Data radio bearer identifier	To identify a DRB per UE	4 bits	Temporary
BI	Linked EPS bearer ID	To identify the default bearer associated with a dedicated EPS bearer	4 bits	Temporary
EID	Tunnel endpoint identifier	To identify the endpoint of a GTP tunnel when the tunnel is established	32 bits	Temporary
NB S1AP E ID	eNB S1AP UE identity	This is a temporary identity used to identify a UE over the S1-MME reference point within the eNB and is unique within the eNB.	32 bits	Temporary
MME S1AP E ID	MME S1AP UE identity	This is a temporary identity used to identify a UE on the S1-MME reference point within the MME and is unique within the MME.	32 bits	Temporary
-RNTI	Cell-radio network temporary identifier	To identify a UE uniquely in a cell	0 × 0001 ~ 0 × FFF3 (16 bits)	Temporary

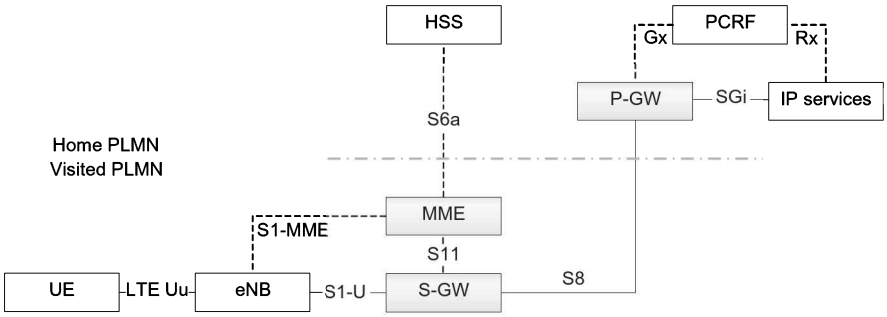


FIGURE 2.11

Example roaming architecture [1,70,71].

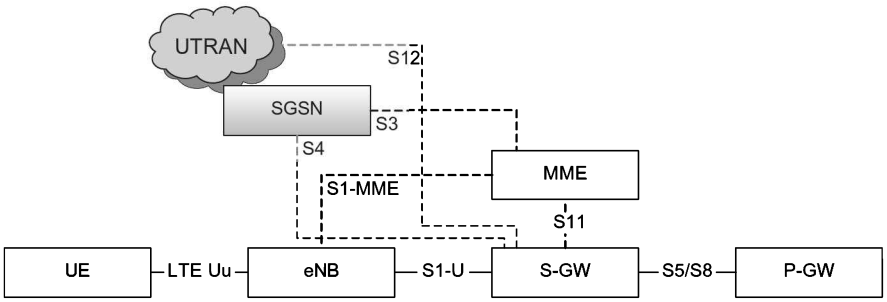


FIGURE 2.12

Interworking of LTE with UMTS network architecture [70,71].

2.4 E-UTRAN architecture

The E-UTRAN access network consists of a network of eNBs, as illustrated in Figure 2.13. For unicast user traffic, there is no centralized controller in the E-UTRAN; thus, the E-UTRAN architecture is considered as a distributed architecture. The eNBs are typically interconnected with each other by means of a logical interface known as X2 and to the EPC by means of the S1 interface. More specifically, the eNB is connected to the MME through S1-MME interface and to the S-GW via S1-U interface. The protocols between the eNBs and the UE are referred to as the AS protocols. The protocols between the EPC and the UE are known as NAS protocols. The E-UTRAN is responsible for all radio-related functions that include RRM, encompassing all functions related to the radio bearers, such as radio bearer control (RBC), radio admission control (RAC), radio mobility control, scheduling and dynamic allocation of radio resources to UEs in both uplink and downlink; header compression, ensuring efficient use of the radio interface by compressing the IP packet headers that could otherwise impose a

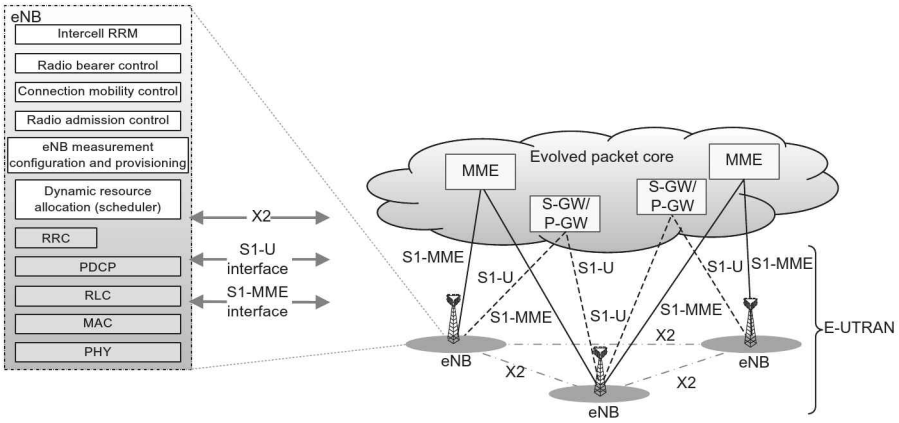


FIGURE 2.13

E-UTRAN architecture and eNB functional decomposition [12].

significant overhead especially for small-packet applications, such as VoIP; security, encrypting all data and control messages sent over the radio interface; and connectivity to the EPC, consisting of the signaling toward MME and the bearer path toward the S-GW [70,71].

As shown in Figure 2.13, on the network side, all of these functions reside in the eNBs, each of which can be responsible for managing multiple cells. Unlike previous 2G/3G technologies, LTE has integrated the radio controller function into the eNB. This allows close interaction between different protocol layers of the radio access network, thus reducing latency and improving efficiency. Such distributed control structure eliminates the need for a processing-intensive central controller, which in turn has the potential to reduce costs, overhead, latency, and complexity. Given that LTE does not support soft handover, there is no need for a centralized data-combining function in the network. One consequence of the absence of a centralized controller node is that, as the UE moves, the network must transfer all information related to a UE (the UE context) together with any buffered data from one eNB to another. Therefore, there are mechanisms in EPS that are provisioned to avoid data loss during handover (e.g., data forwarding).

An important feature of the S1 interface linking the access network to the core network is known as S1-flex [6]. This is a concept whereby multiple core network nodes (MMEs/S-GWs) can serve a common geographical area, being connected by a mesh network to the set of eNBs in that area. An eNB may be served by multiple MMEs/S-GWs. The set of MME/S-GW nodes that serve a common area is called an MME/S-GW pool, and the area covered by such a pool of MMEs/S-GWs is called a pool area. This concept allows the UEs in the cell or cells controlled by one eNB to be shared among multiple core network nodes, thereby providing a possibility for load balancing (LB) and also eliminating a single point

of failure for the core network nodes. The UE context typically remains in the same MME as long as the UE is located within the same pool area.

2.4.1 E-UTRAN functions

The functions performed and the services provided by an eNB include RRM (RBC, RAC, connection mobility control (CMC), dynamic allocation and scheduling of resources for the UEs in the uplink and downlink); IP header compression and encryption of user data stream; selection of an MME upon UE attachment to the network when no routing to an MME can be determined from the information provided by the UE; routing of user-plane data toward S-GW; scheduling and transmission of paging messages originated from the MME; scheduling and transmission of broadcast information originated from the MME or network operation and management; measurement and measurement reporting configuration for mobility and scheduling; scheduling and transmission of emergency warning messages originated from the MME; closed-subscriber group (femto-cell) handling; and transport-level packet marking in the uplink. A relay-enabled eNB further performs S1/X2 proxy functionality for supporting the relay nodes (RNs) and S11 termination and S-GW/P-GW functionality for supporting the relay nodes [12]. The functional decomposition of an eNB is shown in Figure 2.13.

2.4.2 E-UTRAN interfaces

This section provides general information about the E-UTRAN interfaces (S1 and X2). The E-UTRAN network interface model consists of two main parts: the transport network layer, corresponding to the transport of the access-network data, and the radio network layer, which encompasses the top-level protocols of the radio interface. In addition, each interface has a user-plane and a control-plane component. The user-plane carries all user data as well as application layer signaling, such as session initiation protocol (SIP) or real-time transport protocol (RTP) control protocol (RTCP) packets [22]. The control-plane handles all messages and procedures related to the radio interface supported features. An example of these messages includes control messages for handover management or bearer establishment. The physical layer is common to both user-plane and control-plane. The user-plane and control-plane use specific protocols which allow definition of a different and independent transport stack and bearers for each of the planes.

The control-plane information is more susceptible to security, reliability, and information loss, whereas the user-plane information can rely on simpler and less secured routing protocols. The E-UTRAN network interfaces are open standards defined by 3GPP. This conformity to open standards allows different vendors to manufacture eNBs, deploy them in a single network, and interconnect them over the X2 interface or with MME or S-GW nodes over the S1 interface.

The S1 interface connects the eNB to the EPC. It can be split into control-plane (S1-MME) and user-plane (S1-U). The S1-MME is a signaling interface which

supports a set of functions and procedures between the eNB and the MME. All S1-MME signaling procedures belong to four main groups: (1) bearer-level procedures, corresponding to all procedures related to bearer setup, modification and release, which are typically used during the establishment or the release of a session; (2) handover procedures, which encompass all S1 functions related to user mobility between eNB or with 2G/3G technologies; (3) NAS signaling transport, corresponding to the transport of UE-MME NAS signaling over the S1 interface; and (4) paging procedure, which is used in case of a mobile-terminated call/session. Through the paging procedure, the MME requests the eNB to page a UE in a given set of cells known as a tracking area. The S1-MME interface provides a high level of reliability in order to avoid message retransmission and unnecessary delay in control-plane procedures using stream control transmission protocol (SCTP) [20].

The role of S1-U is to transport user data packets between the eNB and the S-GW. This interface makes use of a very simple GTP over user datagram protocol (UDP)/IP transport protocol which only provides user data encapsulation. There is no flow control or error control, or any mechanism to guarantee data delivery over the S1-U interface. The GTP was inherited from GPRS and UMTS networks. In GPRS networks, GTP is used between the SGSN and the GGSN. In UMTS networks, GTP is also used over the Iu-ps interface between the RNC and the SGSN. S1 application part (S1AP) is the control-plane signaling protocol between the eNB and the MME [13]. It supports the S1-MME interface and utilizes SCTP in the transport layer. The SCTP protocol is for the control-plane, which guarantees delivery of control/signaling messages between the MME and the eNB.

The X2 interface connects multiple eNBs. LTE uses the same protocol structure over both S1 and X2 interfaces, which simplifies the data forwarding operation. The X2 interface has a control-plane (X2-C) and a user-plane (X2-U) component. The X2-U is responsible for the transport of user data packets between eNBs. This interface is only used for a limited period of time, when the terminal moves from the coverage area of one eNB to another and provides buffered packet data forwarding. The X2-U interface utilizes GTP tunneling protocol that is used over the S1-U interface. The X2-C is a signaling interface which supports a set of functions and procedures between eNBs. The X2-C procedures are very limited in number and are all related to user mobility between eNBs, exchange of UE context between nodes including allocated bearers, security information, etc. The need for reliable transport of signaling between nodes results in the use of SCTP protocol for X2-C. X2AP is the control-plane protocol between eNBs over the X2 interface and supports load management and handover coordination between eNBs.

2.4.3 E-UTRAN identifiers

The E-UTRAN (dynamically) assigns unique identities to the UEs at the cell level. These identities temporarily identify the UE resource allocations and

assignments as well as UE-specific control signaling during the active state of the UE in the cell. These identities include C-RNTI, which is a unique identification used for identifying RRC connection and scheduling; semi-persistent scheduling C-RNTI that is a unique identification used for semi-persistent scheduling; temporary C-RNTI, TPC-PUSCH-RNTI that is an identification used for the power control of physical uplink shared channel; TPC-PUCCH-RNTI, which is an identification used for the power control of physical uplink control channel; and random access RNTI for random access procedure and contention resolution. During some transient intervals, the UE is temporarily identified with a random value used for contention resolution purposes.

As shown in Figure 2.11, the E-UTRAN further utilizes unique identities for identifying a specific network entity which include GUMMEI that is used to globally identify the MME. The GUMMEI is constructed from the PLMN identity to which the MME belongs, the group identity of the MME group where the MME is part of, and the MMEC of the MME within the MME group. Note that the GUMMEI or S-TMSI containing the MMEC is provided by the UE to the eNB. The E-UTRAN cell global identifier (ECGI) is used to globally identify the cells. The ECGI is constructed from the PLMN identity where the cell belongs to and the cell identity (CI) of the cell. The PLMN information is periodically broadcasted as part of system information. The eNB identifier (eNB ID) is used to identify eNBs within a PLMN and is contained within the CI of its cells. The global eNB ID is used to globally identify the eNBs, and is constructed from the PLMN identity to which the eNB belongs and the eNB ID. The MCC and MNC are the same as that included in the ECGI (see Table 2.2).

In addition to the above identities, there are other identifiers related to the support of home eNBs (HeNBs) or femto-cells and relays. The global eNB ID of a relay node is the same as its serving donor eNB (DeNB), which is a relay-enabled eNB. The tracking area identity is used to identify tracking areas and is constructed from the PLMN identity where the tracking area belongs to, and the TAC of the tracking area. The CSG identity (CSG ID) is an identifier of a CSG femto-cell within a PLMN.

As shown in Figure 2.11, there are unique identifiers that are used to identify the data and signaling radio bearers (SRBs) as well as S1 and S5/S8 bearers on the access network that are assigned to UE services. The value of the E-RAB ID used on S1 and X2 interfaces to identify an E-RAB allocated to the UE is the same as the EPS bearer ID value used at the LTE-Uu interface to identify the associated EPS bearer and the NAS layer. The CI, TAC, CSG ID (if any), and one or more PLMN IDs are broadcast periodically in every E-UTRAN cell as part of system information.

2.4.4 Support of HeNB in E-UTRAN

Small cells are low-power wireless access points that operate in licensed spectrum and are managed by an operator. These cells provide improved cellular coverage

and capacity and have applications for homes and enterprises as well as metropolitan public areas. There are various types of small cells including femto-cells, pico-cells, metro-cells and micro-cells, broadly increasing in size from femto-cells (the smallest) to micro-cells (the largest). Small cells also facilitate a new class of mobile services that exploit the technology's ability to detect the presence of, and connect and interact with existing networks [42,56,60].

As shown in Figure 2.14, the E-UTRAN architecture supports deployment of a home eNB gateway (HeNB-GW) to allow an S1 interface between the HeNB and the EPC in order to support a large number of HeNBs in a scalable manner. The HeNB-GW serves as a concentrator for the control-plane, specifically the S1-MME interface. The S1-U interface from the HeNB may be terminated at the HeNB-GW or a direct logical user-plane connection between HeNB and S-GW may be used. Depending on the deployment scenario, the S1 interface in this case is defined as the interface between the HeNB-GW and the core network; between the HeNB and the HeNB-GW; between the HeNB and the core network; or between the eNB and the core network. The HeNB-GW appears to the MME as an eNB, and to the HeNB as an MME. The S1 interface between the HeNB and the EPC is the same, regardless of the HeNB connection to the EPC via a HeNB-GW [12].

The HeNB-GW connects to the EPC such that the inbound and outbound mobility to cells served by the HeNB-GW does not necessarily require inter-MME handovers. One HeNB serves only one cell. In general, the functions supported by the HeNB are the same as those supported by an eNB and the procedures activated between a HeNB and the EPC are similar to those between an eNB and the EPC. Therefore, it can be concluded that HeNB is a (reduced-function and reduced-cost)

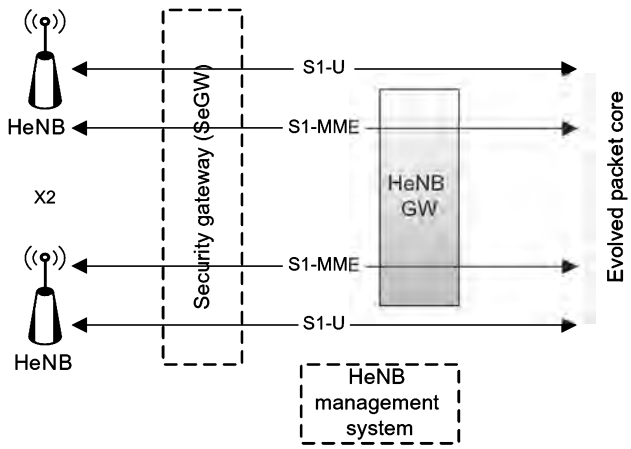


FIGURE 2.14

HeNB logical network architecture [12].

eNB that is installed by the user and not the operator, but is connected to operator's network in the same way that a regular eNB is connected.

3GPP Rel-11 specifications support direct X2-connectivity between HeNBs, independent of whether any of the involved HeNBs are connected to a HeNB-GW [12].

2.4.5 Support of RRHs in E-UTRAN

One recent deployment trend that is foreseen to be widely applied to LTE networks consists of splitting the base station functionalities into a baseband unit, which performs, in particular, the scheduling and the baseband processing functions, and a number of RRHs responsible for all RF transmission/reception aspects. The baseband processing unit is typically located at the center of the cell and is connected via optical fiber to the RRHs. In addition to suppressing the feeder loss (since the power amplifier is immediately close to the antenna), this approach allows a baseband processing unit to manage different radio sites in a central manner. Furthermore, having geographically separated RRHs controlled from the same location enables either centralized baseband processing units jointly managing the operation of several radio sites, or the exchange of very low-latency coordination messages among individual baseband processing units.

Distributed base stations with RRHs can significantly assist mobile operators to overcome cost, performance, and efficiency issues when addressing increasing demand for higher capacities in cellular networks. Multimode radios operating according to GSM, HSPA, LTE, and Wi-Fi standards and advanced software configurability are the key features in the deployment of more flexible and energy-efficient radio networks. Wireless and mobile network operators face the continuing challenge of upgrading their networks to effectively manage high user traffic rates. Mobility and an increased level of multimedia content for end users require end-to-end network adaptations that support both new services and the increased demand for broadband and flat-rate Internet access. In addition, network operators must consider the most cost-effective evolution of the networks toward new generations. Wireless and mobile technology standards are evolving toward higher bandwidth requirements for both peak data rates and cell throughputs. The latest standards HSPA +, mobile WiMAX, and LTE/LTE-Advanced can barely satisfy today's demand for high-performance mobile applications and fast and reliable access to the Internet anywhere and anytime. The network upgrades required to deploy networks based on these standards must balance the limited availability of new spectrum, leverage existing spectrum, and ensure proper operation of the desired wireless standards.

Distributed open base station architecture concepts have evolved in parallel with the evolution of the standards to provide a flexible, low-cost, and scalable modular environment for managing the radio access. For example, the open base

station architecture initiative (OBSAI)[4] and the common public radio interface (CPRI)[5] standards introduced standardized interfaces separating the base station server and the RRH part of a base station by an optical fiber.

The RRH concept constitutes a fundamental part of a state-of-the-art base station architecture. The RRH-based system implementation is driven by the need to reduce both capital expenditure (CAPEX) and operational expenditure (OPEX) consistently, which allows an optimized and energy-efficient deployment. Figure 2.15 illustrates an architecture where 2G/3G/4G base stations are connected to RRHs over optical fibers. Either CPRI or OBSAI may be used to carry RF signals to the RRH to cover a three-sector cell. The RRH incorporates a large number of digital interfacing and processing functions. It also includes high-performance, efficient, and frequency-agile analog functions, all packaged into a light device with a small mechanical footprint.

2.4.6 Support of relays in E-UTRAN

In general, RNs can be used to improve urban or indoor user throughput as well as to extend coverage in rural areas. The concept of relaying is not new but the level of sophistication continues to be enhanced. Figure 2.16 shows a typical scenario where an RN is connected wirelessly to the radio access network via a donor cell. The RN can connect to the Donor cell's eNB (DeNB) in one of two ways: in-band (in-channel), in which case the DeNB-to-RN link shares the same carrier frequency with RN-to-UE links, and out-band, in which case the DeNB-to-RN link does not operate on the same carrier frequency as RN-to-UE links. The most basic relay method is the use of a radio repeater, which receives, amplifies, and then retransmits the downlink and uplink signals to overcome areas of poor coverage. The repeater

[4]The OBSAI family of specifications provides the architecture, functional description, and minimum requirements for integration of a set of common modules into a base transceiver station. It defines an open, standardized internal modular structure of wireless base stations; a set of standard BTS modules with specified form, fit, and function such that BTS vendors can acquire and integrate modules from multiple vendors in an OEM fashion; defines open, standards-based internal digital interfaces between BTS modules to ensure interoperability and compatibility; and supports different access technologies, such as GSM/EDGE, CDMA2000, WCDMA, LTE/LTE-Advanced, or mobile WIMAX [75].

[5]The CPRI is an initiative to define a publicly available specification that standardizes the protocol interface between the radio equipment control (REC) and the radio equipment (RE) in wireless base stations. This allows interoperability of equipments from different vendors and preserves the software investment made by wireless service providers. Conventional base stations are located adjacent to the antenna in a small compartment at the base of the antenna tower. Finding suitable sites can be a challenge because of the footprint required for the compartment, the need for structural reinforcement of rooftops, and the availability of both primary and backup power sources. The CPRI allows the use of a distributed architecture where base stations, containing the REC, are connected to remote radio heads via low-loss fiber links that carry the CPRI data. This architecture reduces costs for service providers because only the remote radio heads containing the RE need to be situated in environmentally challenging locations. The base stations can be centrally located in less problematic locations where footprint, climate, and availability of power are easily managed [38,44].

Samsung Ex. 1005, Page 95 of 1147

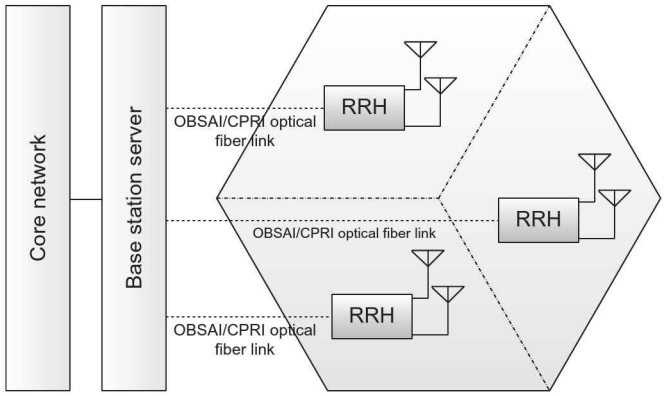


FIGURE 2.15

Example three-sector cell using RRH [38].

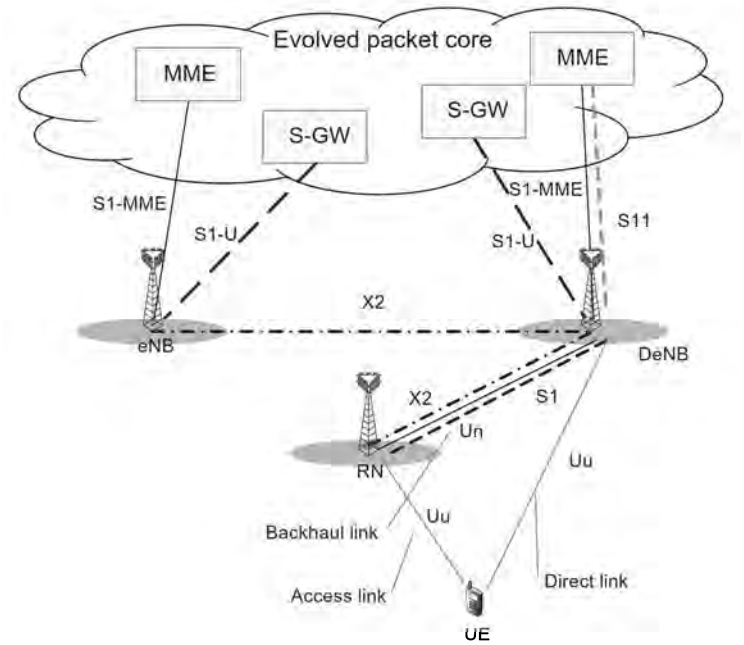


FIGURE 2.16

Support of relay nodes in E-UTRAN architecture [12].

can be located at the cell-edge or in some other areas of poor coverage. Radio repeaters are relatively simple devices operating purely at the RF level. Typically they receive and retransmit an entire frequency band, so they must be sited carefully. In general, repeaters can improve coverage but do not substantially increase capacity. In an in-band relaying scenario and due to self-interference, the RNs cannot simultaneously transmit on the access link and receive on the backhaul link as well as they cannot concurrently receive on the access and transmit on the backhaul link.

E-UTRAN supports relaying functionality by connecting an RN to a serving eNB, known as the donor eNB, via a modified version of the E-UTRA radio interface, where the modified version is denoted as the Un interface. The RN supports most of the eNB functionalities, implying that it terminates the radio protocols of the E-UTRA radio interface on the access link as well as the S1 and X2 interfaces on the backhaul link. In addition to the eNB functionality, the RN also supports a subset of the UE functionality, e.g., physical layer, layer-2, RRC, and NAS functionality, in order to wirelessly connect to the DeNB. An RN may not use another RN as its DeNB and intercell handover of the RN is not supported.

More advanced relays at layer 2 can decode transmissions before retransmitting them. Traffic can then be forwarded selectively to the UE by the RN, thus minimizing the interference created by legacy relays that forwarded all traffic. Depending on the level at which the protocol stack is terminated in the RN, such types of relay may require the development of relay-specific standards. This can be avoided by extending the protocol stack of the RN up to layer 3 to create a wireless router that operates in the same way that a normal eNB operates, using standard air-interface protocols and performing its own resource allocation and scheduling.

The modified E-UTRAN architecture to support RNs is shown in Figure 2.16. The RN terminates the S1, X2, and Un interfaces. The DeNB provides S1 and X2 proxy functionality between the RN and other eNBs, MMEs, and S-GWs. The S1 and X2 proxy functionality includes forwarding UE-specific S1 and X2 signaling messages as well as GTP data packets between the S1 and X2 interfaces associated with the RN and the S1 and X2 interfaces associated with other network nodes. Due to the proxy functionality, the DeNB appears as an MME (for S1-MME), an eNB (for X2) and an S-GW (for S1-U) to the RN. In further development of relay specifications in 3GPP, the DeNB also implements and provides the S-GW/P-GW functions to the associated RNs. This includes creating a session for the RN and managing EPS bearers for the RN, as well as terminating the S11 interface toward the MME serving the RN. The RN and DeNB also map signaling and data packets to the EPS bearers that are set up for the RN. The mapping is based on the existing QoS mechanisms already defined for the UE and the P-GW. The P-GW functions in the DeNB allocate an IP address for the RN for the operation and management functions which may be different than the S1 IP address of the DeNB [12].

The S1 and X2 user-plane packets are mapped to radio bearers over the Un interface. The mapping can be based on the QCI associated with the UE's EPS bearers. The UE's EPS bearers with similar QoS can be mapped to the same Un radio bearer. There is a unique S1 interface relationship between each RN and its DeNB, and

between the DeNB and each MME in the MME pool. The DeNB processes and forwards all S1 messages between the RN and the MMEs for all UE-dedicated procedures. The processing of S1AP messages includes modifying S1AP UE IDs, transport layer address and GTP TEIDs, leaving other parts of the message unchanged [12].

All non-UE-dedicated S1AP procedures are terminated at the DeNB, and are handled locally between the RN and the DeNB, and between the DeNB and the MME(s). Upon reception of an S1 non-UE-dedicated message from an MME, the DeNB may trigger corresponding S1 non-UE-dedicated procedure(s) to the RN(s). If more than one RN is involved (in a multi-hop scenario), the DeNB may wait and aggregate the response messages from all involved RNs before responding to the MME. Upon reception of an S1 non-UE-dedicated message from an RN, the DeNB may trigger associated S1 non-UE-dedicated procedure(s) to the MME(s). There is a single X2 interface relation between each RN and its DeNB. In addition, the DeNB may have X2 interface connection to neighboring eNBs. The DeNB processes and forwards all X2 messages between the RN and other eNBs for all UE-dedicated procedures. The processing of X2AP messages includes modifying S1/X2AP UE IDs, transport layer address, and GTP TEIDs, leaving other parts of the message intact [12].

All non-UE-dedicated X2AP procedures are terminated at the DeNB, and are handled locally between the RN and the DeNB, and between the DeNB and other eNBs. Upon reception of an X2 non-cell-related non-UE-associated message from RN or neighbor eNB, the DeNB may trigger associated non-UE-dedicated X2AP procedure(s) to the neighbor eNB or RN(s). Upon reception of an X2 cell-related non-UE-dedicated message from RN or neighbor eNB, the DeNB may pass associated information to the neighbor eNB or RN(s) based on the included cell information. If one or more RN(s) are involved, the DeNB may wait and aggregate the response messages from all participating nodes to respond to the originating node. The S1 and X2 interface signaling packets are mapped to radio bearers over the Un interface. The RN connects to the DeNB via the Un interface using the same radio protocols and procedures as a UE connecting to an eNB [12].

2.5 E-MBMS architecture

Enhanced multimedia broadcast multicast service (E-MBMS) is a point-to-multi-point service in which data is transmitted from a single source to multiple destinations over the radio network. Transmitting the same data to multiple recipients allows network resources to be shared. E-MBMS is realized by addition of new functional entities to 3GPP network architecture. In practice, E-MBMS provides two different services: broadcast and multicast.

The broadcast service can be received by any subscriber located in the area in which the service is offered and multicast services can only be received by users subscribed to the service and joined to the multicast group associated with the service. Both services are unidirectional point-to-multi-point transmissions of

multimedia data, text, audio, video, news, etc., from a broadcast multicast service center (BM-SC) to any user located in the service area. For such a service, only the broadcast service providers can be charged, possibly based on the amount of broadcast data size of service area, or broadcast service duration. Multicast is subject to service subscription and requires the end user to explicitly join the group in order to receive the service. Because it is subject to subscription, the multicast service allows the operator to set specific user charging rules for this service.

The E-MBMS content is typically broadcast over a large number of cells that transmit the same data and are required to be synchronized. In that case, the resulting signal will appear to a UE as one transmission over a time-disperse radio channel. This concept is known as multimedia broadcast single frequency network (MBSFN). A group of cells that transmit the same content to multiple users form an MBSFN area. Multiple cells can belong to a single MBSFN area and each cell can be part of up to eight MBSFN areas. There can be up to 256 different MBSFN areas defined in a network, each one with a unique identity. Once defined, MBSFN areas will not change dynamically. It is not required that a terminal simultaneously receive content from multiple MBSFN areas. In other words, an MBSFN area is a geographical area comprising one or several cells that transmit the same content. For example, in Figure 2.17, which illustrates a hypothetical MBSFN network, cells 2 and 3 belong to both MBSFN areas 1 and 2. The MBSFN areas are static and do not vary over time. The usage of MBSFN transmission requires not only time synchronization among the cells participating in an MBSFN area, but also usage of the same set of radio resources in each of the cells for a particular service. This coordination is the responsibility of the multi-cell/multicast coordination entity (MCE), which is a logical node in the radio access network that controls resource allocation and transmission parameters across the cells in the MBSFN area [12,30–33].

Figure 2.18 illustrates the E-MBMS network architecture. As shown in the figure, the MCE can control multiple eNBs, each controlling one or more cells. The BM-SC, located in the core network, is responsible for authorization and authentication of content providers, charging, and overall configuration of the data flow through the core network. The MBMS gateway (MBMS-GW) is a logical node handling multicast of IP packets from BM-SC to all eNBs involved in transmission in an MBSFN area. It also manages session control signaling via the MME. From the BM-SC, the MBMS data is forwarded using IP-based multicast, a method of sending IP packets to multiple receiving network nodes in a single transmission, via the MBMS-GW to the cells. The MBMS is not only efficient from a radio-interface point of view, but also it saves resources in the transport network by avoiding unicast transmission of the same packet to multiple nodes unless necessary. This can lead to significant savings in the transport network.

More specifically, an MBMS-enabled network consists of the following entities and interfaces [12,32]:

- **MCE:** The MCE is a logical entity (this does not preclude the possibility that it may be part of another network element), whose functions are admission

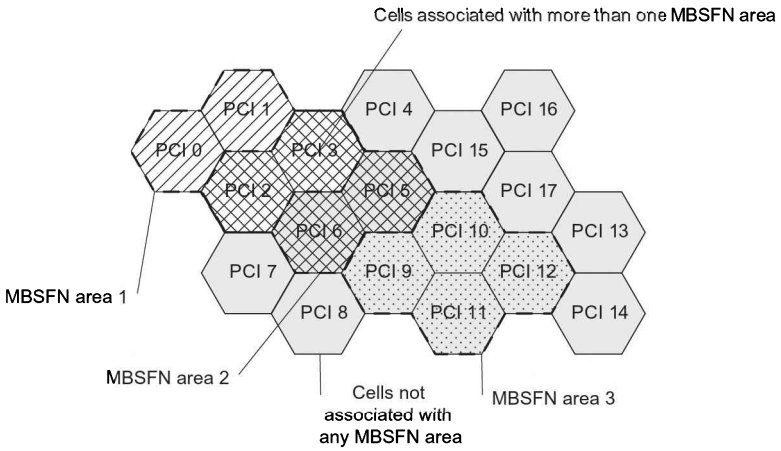


FIGURE 2.17

Illustration of MBSFN areas.

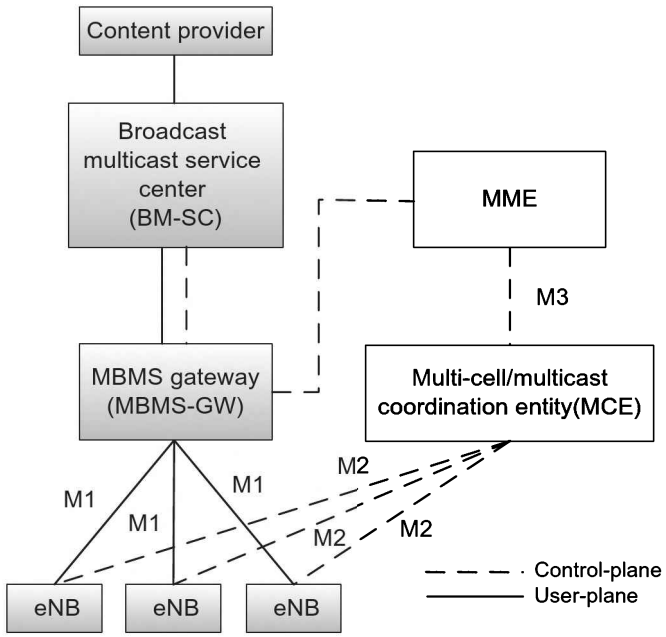


FIGURE 2.18

E-MBMS network architecture [12,32].

control, and allocation of the radio resources used by all eNBs in the MBSFN area for multi-cell MBMS transmissions using MBSFN operation. The MCE decides whether to establish radio bearer(s) for the new MBMS service(s), if the radio resources are not sufficient for the corresponding MBMS service(s), or may allocate radio resources from other radio bearer(s) of ongoing MBMS service(s) according to allocation and retention priority (ARP). In addition to allocation of the time/frequency radio resources, the MCE controls the radio configuration, for example, modulation and coding scheme; counting and acquisition of counting results for MBMS service(s); resumption of MBMS session(s) within MBSFN area(s) based on the ARP and/or the counting results for the corresponding MBMS service(s); suspension of MBMS session(s) within MBSFN area(s) based on the ARP and/or on the counting results for the corresponding MBMS service(s). In the case of distributed MCE architecture, the MCE manages the above functions for a single eNB of an MBSFN.

- **E-MBMS Gateway (MBMS-GW):** The MBMS-GW is a logical entity (this does not preclude the possibility that it may be part of another network element) that is located between the BM-SC and eNBs whose functions include sending/broadcasting of MBMS packets to each eNB participating in the service. The MBMS-GW uses IP multicast as the method of forwarding MBMS user data to the eNBs. The MBMS-GW performs MBMS session control signaling (i.e., session start/update/stop) toward the E-UTRAN via MME.
- **MCE–MME Interface (M3):** The application part of this control-plane interface allows MBMS session control signaling on E-RAB level (i.e., does not convey radio configuration data). The procedures include MBMS session start and stop. The SCTP protocol[6] is used as signaling transport method.
- **MCE–eNB Interface (M2):** The application part of this control-plane interface conveys radio configuration data for the multi-cell transmission mode eNBs and session control signaling. The SCTP protocol is used as signaling transport.
- **MBMS–GW-eNB Interface (M1):** This user-plane interface carries IP packets between the two network entities and uses IP multicast for point-to-multipoint delivery of user packets.

[6]SCTP is a transport layer protocol, performing a similar role as the commonly used protocols TCP and UDP. It provides some of the service features of both protocols, that is, it is message-oriented like UDP and ensures reliable and in-sequence transport of messages with congestion control like TCP. The protocol is defined by IETF RFC 4960 [20]. The SCTP protocol may be characterized as message-oriented, that is, it transports a sequence of messages comprising a group of bytes, rather than transporting an unbroken stream of bytes as does TCP. As in UDP, in SCTP a sender sends a message in one operation, and that exact message is passed to the receiving application process in one operation. In contrast, TCP is a stream-oriented protocol, transporting streams of bytes reliably and in order.

Samsung Ex. 1005, Page 101 of 1147

There are two MBMS architectures defined in 3GPP. In the "distributed MCE architecture," the MCE is part of the eNB and the M2 interface is maintained between the MCE and the corresponding eNB, whereas in the "centralized MCE architecture" (shown in Figure 2.18), the MCE is a logical entity and can be deployed as a stand-alone physical entity or collocated in another physical entity, such as eNB. In both cases, the M2 interface is maintained between the MCE and all eNB(s) belonging to the corresponding MBSFN area [12].

2.6 Positioning network architecture

3GPP legacy technologies (GSM and UMTS) have supported location services network architecture for the purpose of positioning of mobile devices since 3GPP Rel-4. With the introduction of EPS in 3GPP Rel-8, control-plane location-service architecture for the EPS was introduced in 3GPP Rel-9. Figure 2.19 illustrates the control-plane location service architecture for the EPS, which introduced new interfaces SLg and SLs; and allowed the EPS control-plane element MME to interconnect with the location service core network elements, enabling location services using the positioning functionality provided by the E-UTRAN.

The location service (LCS) architecture follows a client/server model with the gateway mobile location center (GMLC) acting as the location server providing location information to LCS clients. The GMLC sends location requests to the evolved serving mobile location center (E-SMLC) through the MME to retrieve this location information. The E-SMLC is responsible for interaction with the UE through E-UTRAN to obtain the UE position estimate or to obtain position measurements that help the E-SMLC to estimate the UE position. Note that the GMLC interaction over the interfaces connected to it other than SLg was already available before 3GPP Rel-9 for GSM and UMTS access.

The LCS clients can be part of the core network or external to the core network and can also reside in the UE or attached to the UE. Depending on the location of the LCS client, the location request initiated by the LCS client may be a mobile-originated, mobile-terminated, or network-induced location request. Emergency location service is possible even if the UE does not have a valid service subscription due to regulatory requirements. Support of location service-related functionality in the E-UTRAN, MME, and UE are optional. The LCS is applicable to any target UE no matter if the UE supports LCS. The following is a description of various LCS architectural elements and interfaces shown in Figure 2.19 [30]:

- **GMLC:** This entity receives and processes location requests from LCS clients; obtains routing information from HSS; performs registration authorization; communicates information needed for authorization, location service requests, and location information with other GMLC; checks the target UE privacy profile settings; and depending on roaming support, it may assume the role of requesting GMLC, visited GMLC, or home GMLC.

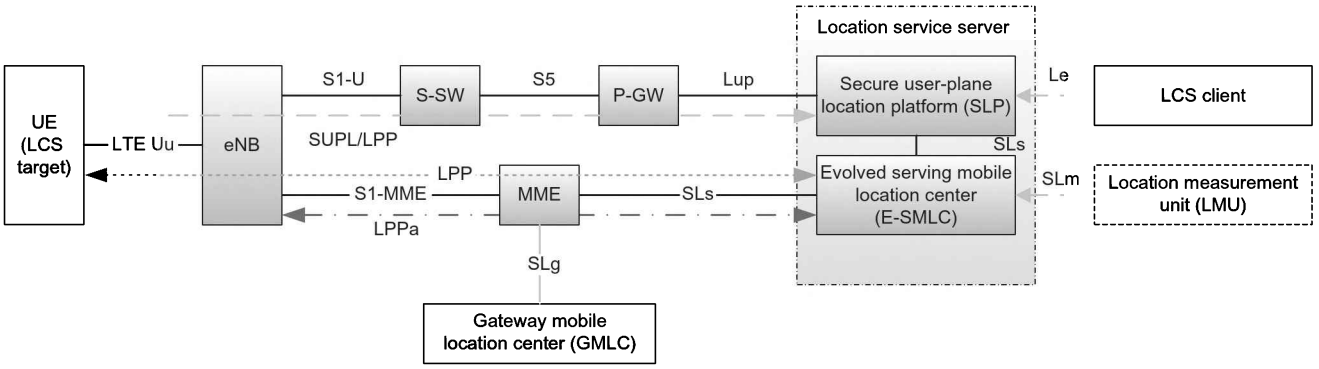


FIGURE 2.19

Positioning network architecture [30].

- **Evolved Serving Mobile Location Center (E-SMLC):** This entity manages the overall coordination and scheduling of resources required for the location determination of a UE that is attached to E-UTRAN. It calculates the final location and estimates velocity as well as the location accuracy. It interacts with the UE to exchange location information applicable to the UE-assisted and the UE-based positioning methods. It interacts with the E-UTRAN to exchange location information applicable to the network-assisted and the network-based positioning methods.
- **Location Measurement Unit (LMU):** The LMU conducts measurements and communicates the measurements to an E-SMLC. All positioning measurements obtained by an LMU are sent to the E-SMLC that requested the measurement report. A UE positioning request may involve measurements by multiple LMUs. The SLm interface between the E-SMLC and the LMU is used for uplink positioning. It is used to transport SLm-AP protocol messages over the E-SMLC-LMU interface. Both stand-alone LMU and LMU integrated into an eNB are supported in LTE Rel-11.

The MME is responsible for UE subscription authorization. It coordinates location services and positioning requests. It handles charging and billing and performs E-SMLC selection (e.g., based on network topology to balance the load on E-SMLC; LCS client type; and requested QoS). It is further responsible for authorization and operation of the LCS services. The HSS is used for storage of location services subscription data and routing information. Two protocols are used for the overall information exchange concerning location services: (1) LTE positioning protocol (LPP); and (2) LTE positioning protocol annex (LPPa). The latter is used for communication between the location server and the eNB. The base station (when using observed time difference of arrival method) is in charge of properly configuring the radio signals that are used by the terminal for positioning measurements, that is, positioning reference signals. It further provides information to the E-SMLC; it enables the device to conduct inter-frequency measurements, if required; and based on the E-SMLC request, it conducts measurements and sends the results to the server. The LPP can be used in user-plane and control-plane methods. The LPP is a point-to-point protocol that allows multiple connections to different devices. The exchanged LPP messages and information can be divided into four categories: (1) UE positioning capability information transfer to the E-SMLC; (2) positioning assisting data delivery from the E-SMLC to the device; (3) location information transfer; and (4) session management.

The SLm-AP protocol, terminated between the E-SMLC and the LMU, is used to support the following functions: (1) delivery of target UE configuration data from the E-SMLC to the LMU; and (2) request positioning measurements from the LMU and delivery of positioning measurements to the E-SMLC. The SLm-AP protocol is directly used between the E-SMLC and the LMU. The MME receives a request for location service associated with a particular target UE from another entity (e.g., GMLC, eNB, or UE), or the MME by itself decides to initiate some

location service on behalf of a particular target UE (e.g., for an IMS emergency call from the UE). The MME then sends a location service request to the E-SMLC. The E-SMLC processes the location service request which may include transferring assisting data to the target UE to help UE-based and/or UE-assisted positioning. The E-SMLC then returns the result of the location service to the MME, for example, a position estimate for the UE and/or an indication of any assisting data transferred to the UE. In the case of a location service requested by an entity other than the MME (e.g., UE, eNB, or E-SMLC), the MME returns the location service result to this entity.

2.7 Mobility management

The NAS is a set of protocols in the EPS. The NAS is used to convey non-radio-access-related signaling between the UE and the MME for an E-UTRAN access. The NAS procedures are grouped in two categories: EPS mobility management (EMM) and ECM. The EMM protocol refers to procedures related to mobility over an E-UTRAN access, authentication, and security. The EMM-specific procedures are UE-initiated. These procedures define attach/detach (to/from the EPC) mechanisms. They also introduce the tracking area update mechanism, which updates the location of the UE within the network. In EPS, a UE initiates a tracking area update when it detects that it enters into a new tracking area. The EMM-specific procedures also define periodic tracking area update. The EMM connection management procedures provide several functions to support the connection of the UE to the network.

The states of a UE with respect to mobility and connection establishment are described by the following NAS states: the EMM state (EMM-DEREGISTERED or EMM-REGISTERED) reflects whether the UE is registered in the MME and the ECM state (ECM-IDLE or ECM-CONNECTED) reflects the connectivity of the UE with the EPC. The transition from ECM-IDLE to ECM-CONNECTED not only involves establishment of an RRC connection, but also includes establishment of an S1 connection. The RRC connection establishment is initiated by the NAS and is completed prior to S1 connection establishment, which means that connectivity in RRC_CONNECTED is initially limited to the exchange of control information between the UE and eNB. The UEs are typically moved to ECM-CONNECTED when becoming active. It should be noted that in LTE, the transition from ECM-IDLE to ECM-CONNECTED is performed within 100 ms and the UEs engaged in intermittent data transfer do not need to stay in ECM-CONNECTED, if the ongoing services can tolerate such transfer delays. The RRC connection release is initiated by the eNB following release of the S1connection between the eNB and the core network [12,31].

The EMM states describe the mobility management states that result from the mobility management procedures, for example, attach and tracking area update procedures. The EMM has two states: EMM-REGISTERED and

EMM-DEREGISTERED. In EMM-DEREGISTERED state, the EMM context does not have any valid location or routing information for the UE. Since the location of the UE is not known, the UE is not reachable by the MME. In order to avoid performing an authentication and key agreement (AKA)[7] security procedure in every network attach attempt by the UE, the MME, and the UE may store some UE context in this state. The MME may perform an implicit detach any time after the implicit detach timer expires. The state is changed to EMM-DEREGISTERED in the MME after performing the implicit detach. If all the bearers belonging to a UE are released (e.g., after a handover from E-UTRAN to non-3GPP access), the MME changes the mobility management state of the UE to EMM-DEREGISTERED. This ensures that the UE performs a tracking area update when it reselects E-UTRAN [12,31].

The ECM states describe the signaling connectivity between the UE and the EPC. In EMM-REGISTERED state, there are two EPS connection states: ECM-IDLE and ECM-CONNECTED. In ECM-IDLE, a UE performs cell selection or reselection. There is no NAS signaling connection between the UE and the network. The UE does not have a context in the E-UTRAN and has no S1-MME, and S1-U connection with the network. In EMM-REGISTERED and ECM-IDLE state, the UE performs a tracking area update, if the current TA is not in the list of TAs that the UE has received from the network in order to maintain the registration and enable the MME to page the UE. The UE performs a periodic tracking area update procedure to notify the EPC that it is available [12,31].

2.8 Radio resource management

The purpose of RRM function is to ensure efficient use of the available network resources. In particular, RRM in E-UTRAN provides a means to manage (e.g., assign, reassign, and release) radio resources in single and multi-cell environments. RRM may be treated as a central application at the eNB responsible for interworking between different protocols (RRC, S1AP, and X2AP) so that messages can be properly transferred to different nodes across Uu, S1, and X2 interfaces. RRM may interface with operation and management functions in order to control, monitor, audit, or reset the status due to errors at a protocol stack. The RRM consists of the following main functions [33,66]:

- Radio admission control: The RAC functional module accepts or rejects requests for the establishment of new radio bearers. Admission control is performed according to the required QoS, current system load, and the required service. The RAC function ensures efficient radio resource utilization

[7]AKA is a mechanism which performs authentication and session key distribution in UMTS networks. AKA is a challenge-response-based mechanism that uses symmetric cryptography. AKA is typically run on a UMTS IP multimedia services identity module (ISIM), which resides on a smart card device that also provides tamper resistant storage of shared secrets. AKA is defined in IETF RFC 3310 [21].

by accepting radio bearer requests as long as radio resources are available, and proper QoS enforcement for current sessions by rejecting radio bearer requests, if they cannot be accommodated. It interacts with the RBC module to perform its functions.

- Radio bearer control: The RBC functional module manages the establishment, maintenance, and release of radio bearers. It configures the radio resources depending on the current resource availability and usage, as well as QoS requirements of new and ongoing sessions due to mobility or other reasons. The RBC function releases radio resources associated with radio bearers at session termination, handover, or other similar cases. The establishment, maintenance, or release of radio bearers is tied to the configuration of radio resources associated with them.
- Connection mobility control: The CMC functional module manages radio resources in the idle and connected modes. In the idle mode, this module defines criteria and algorithms for cell selection, reselection, and location registration that assist the UE in selecting or camping on the best cell. In addition, the eNB broadcasts parameters that configure the UE measurement and reporting procedures. In the connected mode, this module manages the mobility of radio connections without disruption of services. Handover decisions may be based on measurement results reported by the UE, by the eNB, and other parameters configured for each cell. Handover decisions may consider additional information, such as neighbor cell load condition, traffic distribution, transport and hardware resources, and operator-defined policies. Inter-RAT RRM can be one of the subsets of this function responsible for managing the resources in inter-RAT mobility and handovers. In the idle mode, the cell reselection algorithms are controlled by setting of parameters (e.g., thresholds and hysteresis values) that define the best cell and/or determine when the UE should select a new cell. The E-UTRAN broadcasts parameters that configure the UE measurement and reporting procedures.
- Dynamic resource allocation (DRA) or packet scheduling (PS): The task of DRA or PS is to allocate and de-allocate resources (including physical resource blocks) to user- and control-plane packets. The scheduling function typically considers the QoS requirements associated with the radio bearers, the channel quality feedback from the UEs, buffer status, inter-cell/intra-cell interference condition, etc. The DRA function may take into account the restrictions or preferences on some of the available resource blocks or resource-block sets due to inter-cell interference coordination (ICIC) considerations.
- Inter-cell interference coordination: ICIC function manages radio resources such that intercell interference effects are minimized. The ICIC is essentially a multi-cell RRM function that considers resource usage status and the traffic load conditions across multiple cells.
- Load balancing: LB function has the task of handling non-uniform distribution of the traffic load across multiple cells. The purpose of LB function is to control the load distribution in such a manner that radio resources are

efficiently utilized; to maintain the QoS of the current sessions as much as possible; and to minimize the probability of call drops. The LB algorithms may result in handover or cell reselection decisions with the purpose of redistributing traffic from overloaded cells to under-utilized cells.

- Inter-RAT radio resource management: Inter-RAT RRM is primarily concerned with the management of radio resources in connection with inter-RAT mobility, notably inter-RAT handover. During inter-RAT handover, the handover decision may take into account the involved RATs resource availability as well as the UE capabilities and the operator policies. Inter-RAT RRM may also include additional functions for inter-RAT load balancing for the idle mode and connected mode UEs.
- Subscriber profile ID (SPID) for RAT/frequency priority: The RRM function maps SPID parameters received via the S1 interface to a locally defined configuration in order to apply specific RRM strategies (e.g., to define RRC_IDLE mode priorities and control inter-RAT/inter-frequency handover in RRC_CONNECTED mode). The SPID is an index referring to user information, such as mobility profile and service usage profile. The SPID information is UE-specific and applies to all of its radio bearers. The RRM strategy in E-UTRAN may be based on user-specific information.

2.9 Quality of service/quality of experience

The end-to-end quality of service experienced by an end user results from a combination of elements throughout the protocol stack and E-UTRAN/EPC components. Thus, the performance evaluation of the service requires a detailed performance analysis of the entire network, that is, from the UE up to the application server (AS) or remote UE. The QoE is a subjective measurement of the quality experienced by a user when he/she uses a communication service. In other words, the QoE is defined as the overall acceptability of an application or a service as perceived subjectively by the end user. When evaluating the quality of a service, the objective is to optimize the operation of the network from a perspective which is purely based on objective parameters, or to determine the quality that the user is actually experiencing and its satisfaction level. The QoE takes into account the satisfaction of a user in terms of both content and application performance. In this sense, the introduction of smartphones has been a quantitative leap in user QoE expectations [43,67,68].

The QoE has been traditionally evaluated through subjective tests carried out with the users in order to assess their satisfaction with a mean opinion score value. This approach can be quite expensive, time-consuming, and cannot be used for real-time decision-making to improve the QoE. New methods have been emerged in recent years to estimate the QoE based on certain performance indicators associated with services. A possible solution to instantaneously evaluate the

QoE is to integrate the QoE analysis tools in the mobile terminal itself. If mobile terminals are able to report the measurements to a central server, the QoE assessment process is significantly simplified. Other solutions are focused on including new network elements (e.g., network analyzers and deep packet inspectors) that are responsible for capturing the traffic from a certain service and analyzing its performance. In some cases, the evaluation of video-streaming quality in the mobile terminals is addressed by monitoring objective parameters, such as packet loss rate or jitter.

As mentioned earlier, the E-UTRAN consists of a network of eNBs which are typically interconnected by means of an X2 interface and to the EPC via an S1 interface. The eNB entity plays a critical role in maintaining the end-to-end QoS. The eNB performs QoS-related functions which include admission control and resource reservation; scheduling and distribution of radio resources among the established bearers; and L1/L2 protocol configuration in accordance with the QoS attributes associated with the bearer. The LTE-enabled UEs may support multiple applications at the same time, each having different QoS requirements. This is achieved by establishing different EPS bearers for each QoS flow. The EPS bearers can be classified into two categories based on the characteristics of the QoS they provide: GBR bearers in which resources are permanently allocated and non-GBR bearers which do not guarantee any particular bit rate. In the access network, it is the eNB's responsibility to ensure that the QoS requirements for a particular bearer over the radio interface are met. Each bearer has an associated QCI, characterized by priority, packet delay budget, and admissible packet loss rate; and an ARP used for call admission control. The IP packets mapped to the same EPS bearer receive the same bearer level packet forwarding treatment (e.g., scheduling policy, queue management policy, and rate shaping policy). Thus, the UE is not only responsible for requesting the establishment of EPS bearers for each QoS flow, but also for performing packet filtering in the uplink into different bearers based on TFTs, as P-GW does for the downlink (see Figures 2.4 and 2.5).

It was mentioned earlier that one EPS bearer/E-RAB is established when the UE connects to a PDN, and that remains established throughout the lifetime of the PDN connection to provide the UE with always-on IP connectivity to that PDN. That bearer is referred to as the default bearer. Any additional EPS bearer/E-RAB that is established to the same PDN is referred to as a dedicated bearer. The default bearer QoS parameters are assigned by the network based upon subscription data. The decision to establish or modify a dedicated bearer can only be taken by the EPC and the bearer level QoS parameter values are always assigned by the EPC.

2.10 Security in E-UTRAN

Security over the E-UTRAN air-interface is provided through cryptographic techniques. The backhaul link from the eNB to the core network uses Internet key

exchange (IKE)[8] and the IP security protocol (IPsec)[9] when cryptographic protection is needed. These cryptographic techniques provide end-to-end protection for signaling between the core network and UE. Therefore, the main location where user traffic is exposed to security threats is in the eNB. Moreover, to minimize susceptibility to attacks, the eNB must provide a secure environment that supports the execution of sensitive operations, such as the encryption or decryption of user data and the storage of sensitive data, such as keys for securing UE communication, long-term cryptographic secrets, and vital configuration data. Similarly, the use of sensitive data must be confined to this secure environment. Even with the above security measures in place, one must consider attacks on an eNB, because, if successful, they could give attackers full control of the eNB and its signaling to UEs and other nodes. To limit the effect of a successful attack on one eNB, attackers must not be able to intercept or manipulate user- and control-plane traffic that traverses another eNB, for example, following a handover.

The E-UTRAN and EPS security model is shown in Figure 2.20. The subscriber authentication function in EPS is based on the UMTS AKA protocol. It provides mutual authentication between the UE and core network, ensuring robust charging and guaranteeing that no unauthorized and non-authenticated entities can pose as a valid network node. The EPS AKA provides a root key from which a key hierarchy is derived. The keys in this hierarchy are used to protect signaling and user-plane traffic between the UE and the network. The key hierarchy is derived using cryptographic functions. Furthermore, keys are bound to where, how, and for which purpose they are used. This ensures that keys used for one access network cannot be used in another access network, and that the same key is not used for multiple purposes or with different algorithms. The key hierarchy and bindings also make it possible to routinely and efficiently change the keys used between a UE and eNBs (e.g., during handover) without changing the root key or the keys used to protect signaling between the UE and the core network. For radio-specific signaling, LTE provides integrity, replay protection, and encryption between the UE and the eNB.

The IKE/IPsec protocols can protect the backhaul signaling between the eNB and MME. In addition, LTE-specific protocols provide end-to-end protection of

[8]IKE or IKEv2 is the protocol used to set up a security association in the IPsec protocol suite. The IKE protocol is based on a key-agreement protocol and the Internet security association and key management protocol which is a protocol defined by IETF RFC 2408 for establishing security associations and cryptographic keys in an Internet environment. The IKE uses X.509 certificates for authentication which are either pre-shared or distributed to set up a shared session secret from which cryptographic keys are derived. In addition, a security policy for every peer which will connect must be manually maintained [76].

[9]IPsec is a set of protocols for securing IP-based communications by authenticating and encrypting each IP packet of a data stream. IPs also includes protocols for establishing mutual authentication between agents at the beginning of the session and negotiation of cryptographic keys to be used during the session. IPs can be used to protect data flows between a pair of hosts, between a pair of security gateways, or between a security gateway and a host [24,25,52].

Samsung Ex. 1005, Page 110 of 1147

signaling between the MME and UE. For user-plane traffic, IKE/IPsec can similarly protect the backhaul from the eNB to the S-GW. Support for integrity, replay protection, and encryption is mandatory in the eNB. The user-plane traffic between the UE and eNB is only protected by encryption, as integrity protection would result in excessive overhead. Nevertheless, it is not possible to intelligently inject traffic on behalf of another user: attackers are essentially blind in the sense that any traffic they try to inject would almost certainly decrypt to invalid information.

2.11 Support of voice and messaging services in LTE/LTE-Advanced

2.11.1 IMS architecture

Voice services in LTE are enabled using the IMS which provides a framework for supporting IP-based services and requires new IMS-specific network elements as part of the dedicated core network architecture. The first version of IMS was introduced in 3GPP Rel-5 specifications; however, many enhancements have been made to the baseline in subsequent releases. During the development of the LTE standards, it was assumed that IMS would be commercially available when the LTE networks were deployed, since voice support in LTE relied on the IMS. However, rollout of IMS infrastructure was slower than expected, making voice support in LTE a real challenge for many network operators. A closely related feature to the support of voice is the support of SMS as another key circuit-switched service which was not supported in LTE. In general, the support of voice and messaging services in LTE using IMS remains one of the major targets for the industry. An industry initiative called voice over LTE (VoLTE) was formed in 2010 by the Global System for Mobile Communications Association (GSMA) which has developed a framework for support of voice and SMS over IMS in LTE, including roaming and interconnection issues [53].

The IMS is an access-independent IP connectivity and service control architecture. It provides the framework for IP-based multimedia services in a mobile network and is the most efficient choice to provide VoIP services in LTE. In this section, we focus on voice and messaging services over IMS. A simplified IMS reference architecture is shown in Figure 2.21, which includes interfaces toward legacy and IP-based networks. This architecture provides entities for session management and routing, service support, databases, and interworking.

The IP connectivity access network (IP-CAN) in the case of LTE would consist of the EPS and the E-UTRAN. The call session control function (CSCF)

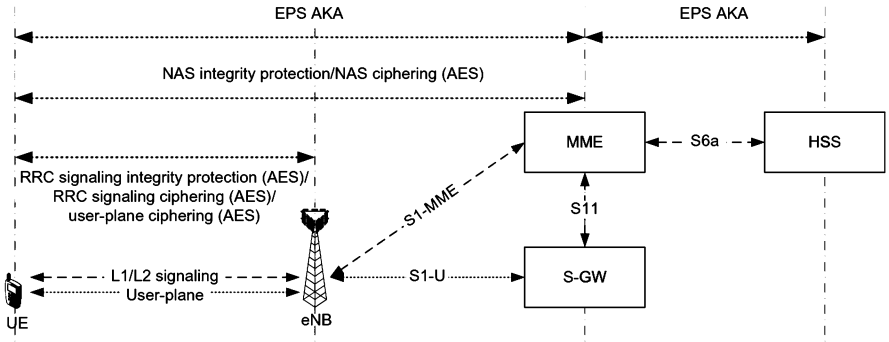


FIGURE 2.20

E-UTRAN and EPS security model [70,71].

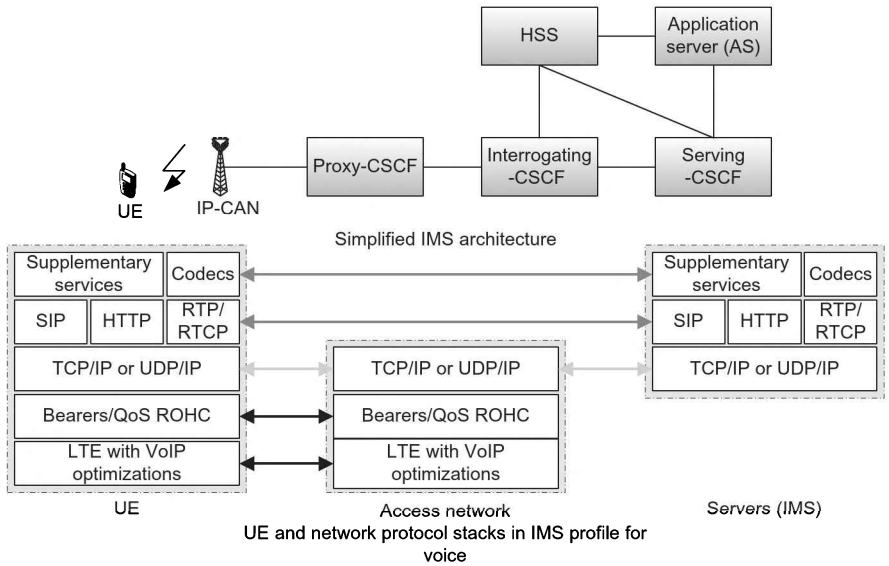


FIGURE 2.21

Simplified IMS architecture and protocol stack [50,65].

entities are the core components of the IMS. There are three CSCF entities defined in IMS architecture:

- Proxy-CSCF (P-CSCF): The P-CSCF is the first point of contact for a user. The P-CSCF behaves as a proxy, that is, it accepts and forwards the requests.
- Interrogating-CSCF (I-CSCF): The I-CSCF is the contact point within an operator's network for all connections destined to a subscriber.

- Serving-CSCF (S-CSCF): The S-CSCF is responsible for handling the registration process, making routing decisions, maintaining sessions, and downloading user information and service profiles from the HSS.

The HSS is the master database for a user. It is comparable to the HLR in a legacy mobile radio network. The HSS contains the subscription-related information required for other network entities that handle calls or sessions. For example, the HSS provides support to the call control servers to complete the routing/roaming procedures by resolving authentication, authorization, naming/addressing resolution, and location dependencies. The application server provides specific IP applications, such as messaging. The purpose of the IMS architecture and the different CSCF entities becomes clearer in the case of roaming. In general, the network operators' internal network structures and user databases are proprietary. The local P-CSCF in the network, which is the point of access for the UE, does not have direct connection to the HSS, rather, it is connected to the HSS via the I-CSCF entity.

The IMS uses a set of IP-based protocols including SIP as a text-based protocol for registration, subscription, notification and initiation of sessions; session description protocol (SDP) as a text-based protocol for negotiating session parameters, such as media type, codec type, bandwidth, IP address and ports, and for media stream setup; RTP and RTCP for transport of real-time applications, such as audio; extensible markup language (XML) configuration access protocol (XCAP), which allows a client to read, write, and modify application configuration data, stored in XML format on a server and maps XML document sub-trees and element attributes to HTTP uniform resource identifiers so that these components can be directly accessed by HTTP; and DHCP for configuring IP addresses for IPv4 and IPv6 [19]. The UDP for messages smaller than 1300 bytes or transport control protocol (TCP) can be used as transport protocols.

As shown in Figure 2.21, the supplementary services requested by the VoLTE profile have been specified in GSMA IR.92 [50,65]. For example, the originating identification presentation (OIP) service provides the terminating user a way to receive identity information to identify the originating user. Another example would be the supplementary service communication hold (HOLD) that enables a user to suspend the reception of media stream(s) from an established IP multimedia session, and resume them at a later time. For supplementary service configuration, the UE and the network must support the XCAP, which is an application layer protocol that allows a client to read, write, and modify application configuration data stored in XML format on a server.

The IMS architecture has been designed to enable operators to provide a wide range of real-time, packet-based services and to track their use in a way that allows both traditional time-based charging as well as packet and service-based charging. It has become increasingly popular both with wire-line and wireless service providers as it is designed to increase carrier revenues, deliver integrated multimedia services, and create an open, standards-based network. The IMS provides a wide range of session border control, including call access control, reachability and security. It also

provides a framework for the deployment of both basic calling services and enhanced services, including multimedia messaging; web integration; presence-based services; and push-to-talk. At the same time, it draws on the traditional telecommunications experience of guaranteed QoS; flexible charging mechanisms (time-based, call-collect, and premium rates); and lawful intercept compliance [50,65].

The IMS decomposes the network infrastructure into separate functions with standardized interfaces between them. Each interface is specified as a "reference point," which defines both the protocol over the interface and the functions between which it operates. The standards do not mandate which functions should be collocated, as this depends on the scale of the application, and a single device may contain several functions. The IMS architecture specified by 3GPP can be divided into three main layers, each of which is described by a number of equivalent names: service/application layer; control layer; and transport layer [8].

Application Layer: An application layer consists of application servers that host the IMS services and an HSS. The application layer provides a framework for provisioning and management of services and defines standard interfaces to common functionalities including configuration storage; identity management; user status, such as presence and location, which is held by the HSS billing services and provided by a charging gateway function (CGF) control of voice and video calls and messaging, provided by the control layer.

Control Layer: The control layer is located between the application and transport layers. It routes the call signaling, informs the transport layer of permissible traffic, and generates billing information for the use of the network. At the core of this layer is the CSCF, which comprises the following functions. The P-CSCF is the first point of contact for users with the IMS. The P-CSCF is responsible for security of the messages between the network and the user and allocating resources for the media flows. The I-CSCF is the first point of contact from peered networks. The I-CSCF is responsible for querying the HSS to determine the S-CSCF for a user and may also hide the operator's inner network from peer networks. The S-CSCF is the central entity which is responsible for processing registrations to record the location of each user, user authentication, and call processing including routing of calls to applications. The operation of the S-CSCF is controlled by policy stored in the HSS. This distributed architecture provides an extremely flexible and scalable solution. For example, any of the CSCF functions can generate billing information for each operation. The control layer also controls transport-layer traffic through the resource and admission control subsystem (RACS). This consists of the PDF, which implements local policy on resource usage, for example, to prevent overload of particular access links, and access-RAC function (A-RACF), which controls QoS within the access network.

Transport Layer: The transport layer provides a core QoS-enabled IPv6 network with access from the UE over Wi-Fi and other mobile broadband networks. This infrastructure is designed to provide a wide range of IP multimedia server-based and P2P services. Access into the core network is through border gateways (GGSN/Packet Data Gateway/Broadband Access Server). These entities enforce

policy provided by the IMS core network, controlling traffic flows between the access and core networks. Within the transport layer, the interconnect border control function (I-BCF) controls transport level security and informs the RACS of the resources required for a call the interconnect border gateway function (I-BGF), access network inter-connect border gateway function (A-BGF) border gateway functions provide media relay for hiding endpoint addresses.

2.11.2 Single radio voice call continuity

Single radio voice call continuity (SRVCC) is an LTE feature that allows a VoIP/IMS call in the LTE packet-switched domain to be transferred to a legacy circuit-switched domain (GSM/UMTS or CDMA2000). Let us assume that a new LTE network operator is going to migrate its voice services to VoIP over IMS in conjunction with deployment of an LTE radio access network. Without SRVCC features, this operator would need to provide a ubiquitous LTE coverage in order to have a viable VoIP service. However, an SRVCC-enabled LTE network may not require comprehensive LTE coverage in order to provide voice services across the entire network. The SRVCC feature provides the ability to transition a voice call from the VoIP/IMS packet-switched domain to the legacy circuit-switched domain. Variations of SRVCC have been standardized to support both GSM/UMTS and CDMA2000 circuit-switched domains. For an operator with a legacy cellular network, deployment of IMS-based VoIP services in conjunction with the rollout of an LTE network is possible through use of SRVCC, which offers VoIP subscribers coverage over a much larger area than would typically be available during the rollout of a new network [3,48,54,72,76].

The SRVCC feature functions as follows. When an SRVCC-capable UE engaged in a voice call determines that it is moving away from an LTE coverage area (based on neighboring cell measurements), it notifies the serving eNB. The LTE network determines that the voice call has to be moved to the legacy circuit-switched domain, thus, it notifies the MSC server of the need to switch the voice call from the packet-switched to the circuit-switched domain and initiates handover of the LTE voice bearer to the circuit-switched network. The MSC server establishes a bearer path for the mobile in the legacy network and notifies the IMS core network that the UE's call is moving from the packet-switched to the circuit-switched domain. The IMS core network then enacts the necessary interworking functions to facilitate the call transfer from one domain to another. When the mobile arrives in the legacy network, it switches its internal voice processing from VoIP to legacy circuit-switched voice processing and the call is resumed. If the legacy circuit-switched network also has an associated packet-switched capability and is capable of supporting concurrent circuit-switched/packet-switched operations, the subscriber's data sessions can be handed over to the legacy network in conjunction with switching the voice call from the packet-switched to the circuit-switched domain. In this case, when the voice call is terminated and the mobile reenters the LTE coverage, these packet sessions can be handed over to the LTE network again.

The SRVCC feature may not be suitable for the operators that do not plan to migrate their voice service to VoIP. If an operator does plan to migrate to VoIP and also plans to roll out ubiquitous LTE coverage, then the question of whether or not to adopt SRVCC is more complicated. While SRVCC does not require modifications to the existing radio access network, it does require a significant modification of the operator's legacy core network and further requires full deployment of IMS circuit-switched/packet-switched continuity services. Given the cost of these changes, deployment of SRVCC as an interim measure to allow early rollout of VoIP-based services may not make financial sense.

The SRVCC provides voice call continuity between LTE radio access and circuit-switched mode of other radio access technologies (e.g., CDMA2000, GSM, or UMTS) using IMS. Calls remain anchored in IMS, which means that the circuit-switched connections can be treated as standard IMS sessions and the circuit-switched bearer is used as a media for IMS sessions. Using SRVCC, the UE only has to be capable of transmitting/receiving on one access network at a given time. The SRVCC feature provides a solution for LTE networks which do not support circuit-switched voice and messaging services. Circuit-switched legacy technologies complement the LTE coverage with voice using IMS support [72].

The SRVCC functionality is built on the IMS centralized services framework for delivering voice and messaging services to the users regardless of the type of network to which they are attached, and for maintaining service continuity for moving terminals. The call anchoring is done by means of a voice call continuity application server in the home IMS network. During EPS attach, routing information to the voice call continuity application server is provided from the HSS to the MME. During the EPS attach procedure, the network and the UE indicate their support for SRVCC. Some modifications in the core network elements are required to support SRVCC, for example, the CDMA2000 circuit-switched SRVCC interworking function is required [18].

To support GSM and UMTS, some modifications in the MSC server are required. When the E-UTRAN selects a target cell for SRVCC handover, it needs to indicate to the MME that this handover procedure requires SRVCC. Upon receiving the handover request, the MME triggers the SRVCC procedure with the MSC server. The MSC then initiates the session transfer procedure to IMS and coordinates it with the circuit-switched handover procedure to the target cell. Handling of any non-voice packet-switched bearer is by the packet-switched bearer splitting function in the MME. The handover of non-voice packet-switched bearers, if performed, is according to a regular inter-RAT packet-switched handover procedure. The handover procedure from E-UTRAN (as shown in Figure 2.22) is used to initiate the SRVCC procedure. When SRVCC is enacted, the downlink flow of voice packets is switched toward the target circuit-switched network. The call is moved from the packet-switched to the circuit-switched domain, and the UE switches from VoIP to circuit-switched voice. Figure 2.23 shows the simplified procedure for SRVCC for GERAN or UTRAN.

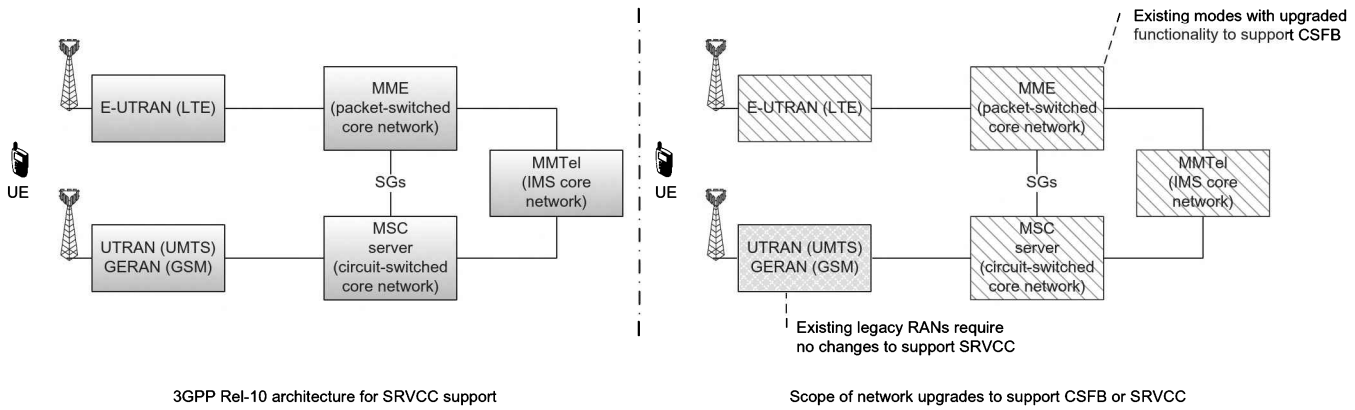


FIGURE 2.22

Network architecture with SRVCC support [72].

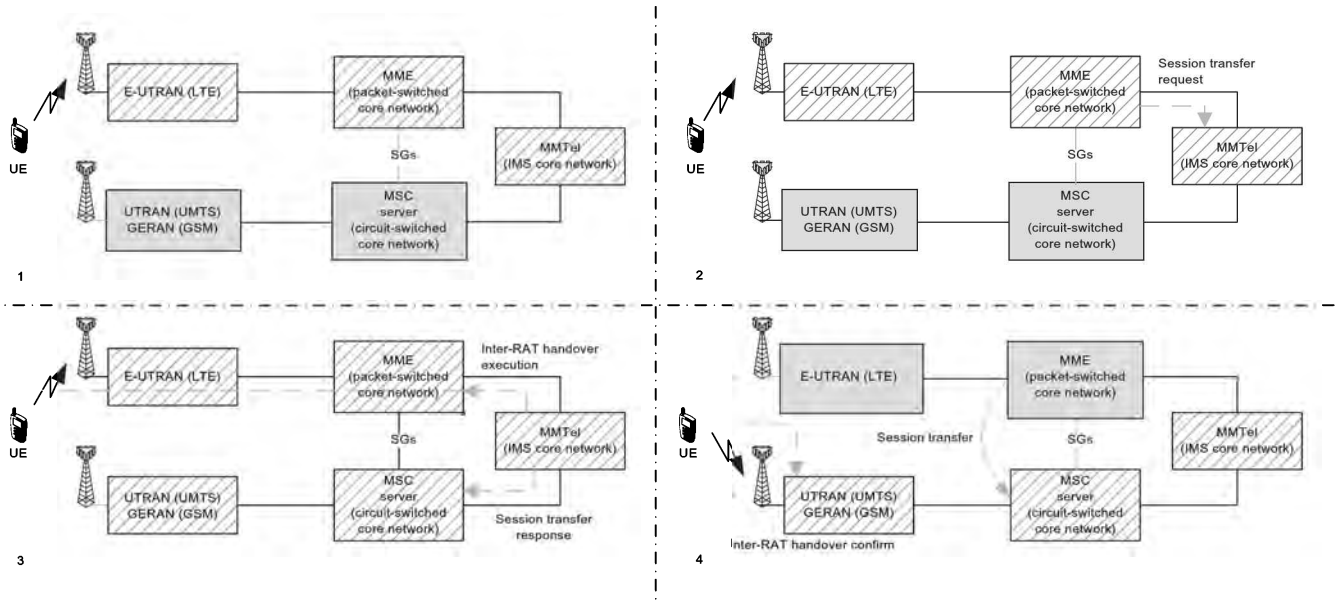


FIGURE 2.23

Call transfer procedure using SRVCC [72].

3GPP Rel-10 architecture has been recommended by GSMA for SRVCC because it reduces both voice interruption time during handover and the dropped call rate compared to earlier configurations. The network controls and moves the UE from E-UTRAN to UTRAN/GERAN as the user moves out of the LTE network coverage area. The SRVCC handover mechanism is entirely network-controlled and calls remain under the control of the IMS core network, which maintains access to subscribed services implemented in the IMS service engine throughout the handover process. 3GPP Rel-10 configuration includes all components needed to manage the time-critical signaling between the user's device and the network, and between network elements within the serving network, including visited networks during roaming. As a result, signaling follows the shortest possible path and is as robust as possible, minimizing voice interruption time caused by switching from the packet-switched core network to the circuit-switched core network, whether the UE is in its home network or roaming. With the industry aligned around the 3GPP standard and GSMA recommendations, SRVCC-enabled user devices and networks will be interoperable, ensuring that solutions work in many scenarios of interest.

As shown in Figure 2.22, the SRVCC functionality can be added to the network by upgrading the mobile soft-switch software as well as the IMS and the LTE/EPC subsystems. The legacy UTRAN/GERAN access networks are not required to be upgraded. When a user with an active voice call using VoLTE moves to an area without LTE coverage, the handover of the call to the network is initiated with inter-RAT handover and is completed with a session transfer. Inter-RAT handover is the conventional method of transferring the UE from E-UTRAN to UTRAN/GERAN. Session transfer is a new mechanism to move access control and voice media anchoring from the EPC to the legacy circuit-switched core network. During the entire voice handover process, the IMS retains control of the user. The handover process is initiated by a session transfer request to the IMS. The IMS/multimedia telephony (MMTel) responds simultaneously with two commands, one to the LTE network and one to the 2G/3G network. The LTE network in which the user's voice call is in progress receives an inter-RAT handover execution command through the MME and the E-UTRAN to instruct the UE to prepare to move to the circuit-switched network. The circuit-switched network, where the user's voice call is being transferred, receives a session transfer response, preparing it to accept the call in progress. With acknowledgments that the commands have been executed, the UE and the IMS, switch to the circuit-switched network to continue the call.

While the implementation of SRVCC does not directly impact the legacy UTRAN/GERAN nodes, an indirect impact arises from methods to help the UE return to the LTE network as soon as the voice call has ended. To help transfer the device back to the LTE network, the legacy RAN can implement functionality to either broadcast LTE system information to the UE, with which the device can perform a cell reselection to an LTE cell after connection release or

release connection to the UE and simultaneously redirect the UE to the LTE network.

The inter-RAT handover and session transfer procedures, as well as the number of sub-systems requiring upgrades, have raised concerns about whether SRVCC can meet the voice interruption time of less than 300 ms. It must be noted that the inter-RAT handover in the E-UTRAN and the session transfer in the core network contribute to an interruption time as they break and make the connection. To address this issue, the SRVCC in 3GPP Rel-10 network architecture has minimized the interruption time by initiating the two procedures simultaneously so that they can be executed in parallel. The studies have shown that the session transfer process can be performed in the order of 10 ms; therefore, the voice call interruption time is mainly affected by the inter-RAT handover latency. The inter-RAT handover delay is measured from the time that the UE receives a network handover command to the time that the UE is uplink synchronized with the new radio access network.

Along with the introduction of the LTE radio access network, 3GPP also standardized SRVCC in Rel-8 specifications to provide seamless service continuity when a UE performs a handover from the E-UTRAN to UTRAN/GERAN. With SRVCC, calls are anchored in the IMS network while the UE is capable of transmitting/receiving on only one of those access networks at a given time, where a call anchored in the IMS core can continue in UMTS/GSM networks and outside of the LTE coverage area. Since its introduction in Rel-8, the SRVCC has evolved with each new release, a brief summary of SRVCC capability and enhancements are noted below [48,49].

- **3GPP Rel-8:** Introduces SRVCC for voice calls that are anchored in the IMS core network from E-UTRAN to CDMA2000 and from E-UTRAN/UTRAN (HSPA) to UTRAN/GERAN circuit-switched. To support this functionality, 3GPP introduced new protocol interface and procedures between MME and MSC for SRVCC from E-UTRAN to UTRAN/GERAN, between SGSN and MSC for SRVCC from UTRAN (HSPA) to UTRAN/GERAN, and between the MME and a 3GPP2-defined interworking function for SRVCC from E-UTRAN to CDMA 2000.
- **3GPP Rel-9:** Introduces the SRVCC support for emergency calls that are anchored in the IMS core network. IMS emergency calls, placed via LTE access, need to continue when SRVCC handover occurs from the LTE network to GSM/UMTS/CDMA2000 networks. This evolution resolves a key regulatory exception. This enhancement supports IMS emergency call continuity from E-UTRAN to CDMA2000 and from E-UTRAN/UTRAN (HSPA) to UTRAN/GERAN circuit-switched network. Functional and interface evolution of EPS entities were needed to support IMS emergency calls with SRVCC.
- **3GPP Rel-10:** Introduces procedures of enhanced SRVCC including support of mid-call feature during SRVCC handover (eSRVCC); support of SRVCC packet-switched to circuit-switched transfer of a call in alerting phase

(aSRVCC); MSC server-assisted mid-call feature enables packet-switched/circuit-switched access transfer for the UEs not using IMS centralized service capabilities, while preserving the provision of mid-call services (inactive sessions or sessions using the conference service). The SRVCC in alerting phase feature adds the ability to perform access transfer of media of an instant message session in packet-switched to circuit-switched direction in alerting phase for access transfers.

- **3GPP Rel-11:** Introduces two new capabilities: single radio video call continuity for 3G-circuit-switched network (vSRVCC); and SRVCC from UTRAN/GERAN to E-UTRAN/HSPA (rSRVCC). The vSRVCC feature provides support of video call handover from E-UTRAN to UTRAN-circuit-switched network for service continuity when the video call is anchored in IMS and the UE is capable of transmitting/receiving on only one of those access networks at a given time. Service continuity from UTRAN/GERAN circuit-switched access to E-UTRAN/HSPA was not specified in 3GPP Rel-8/9/10. To overcome this drawback, 3GPP Rel-11 provided support of voice call continuity from UTRAN/GERAN to E-UTRAN/HSPA. To enable video call transfer from E-UTRAN to UTRAN-circuit-switched network, IMS/EPC is evolved to pass relevant information to the EPC side and S5/S11/Sv/Gx/Gxx interfaces are enhanced for video bearer-related information transfer. To support SRVCC from GERAN to E-UTRAN/HSPA, GERAN specifications are evolved to enable a mobile station and base station sub-system to support seamless service continuity when a mobile station hands over from GERAN circuit-switched access to E-UTRAN/HSPA for a voice call. To support SRVCC from UTRAN to E-UTRAN/HSPA, UTRAN specifications are evolved to enable the RNC to perform rSRVCC handover and to provide relative UE capability information to the RNC.

2.11.3 Voice over LTE

Some of the earliest LTE deployments addressed the lack of circuit-switched voice support in LTE by utilizing a dual-radio solution in which the mobile device was required to simultaneously connect to the E-UTRAN and a legacy circuit-switched network, the latter being used to provide voice and messaging services. The advantage of a simultaneous voice and LTE (SVLTE) solution was that circuit-switched voice and packet-switched data services were supported at the same time. However, an SVLTE-based solution would add significant hardware complexity to the mobile device, given that two completely independent baseband and RF chains were required with obvious cost implications. Furthermore, an SVLTE-based solution has proven to be very power-consuming in practice. On the other hand, since the LTE and the legacy cells do not necessarily have identical footprints, the user experience during mobility can be less than desired. There are several options when it comes to solving the LTE voice services problem. Those range from multi-radio solutions, to solutions which tunnel legacy signaling through LTE, to SIP-based

VoIP solutions, to not using LTE for voice at all, instead relying on legacy networks to provide voice services. The VoLTE solution has developed a framework for optimal support of voice and SMS over IMS in E-UTRAN including roaming and inter-operator issues. VoLTE is based on the existing IMS MMTel[10] concepts and has been specified in GSMA profile IR.92 [45–47]. The GSMA IR.92 (IMS profile for voice and SMS) is a profile as opposed to a technical specification. Therefore, rather than specifying new technologies, this document defines how the framework and services already specified by 3GPP should be configured and utilized in a manner that facilitates interoperability among mobile devices and the serving networks, as well as amongst the networks [64,72,76].

The IR.92 profile requires the mobile device to indicate its support for MMTel-based supplementary services during SIP registration, and if the device supports SMS-over-IP it is also required to be indicated. The terminals are further required to provide their IMEI during SIP registration, which allows the operators to implement management policies in the P-CSCF which are device or device class specific. IR.92-compliant devices are required to support network-initiated deregistration, so that the network can force the mobile device to detach in special circumstances.

IR.92 leverages the existing mechanisms for IMS authentication, that is, IMS AKA. VoLTE-enabled UEs are expected to install a universal integrated circuit card (UICC) equipped with the ISIM application. The ISIM application contains the credentials critical to authentication, such as the subscriber's IP multimedia private identity (IMPI), the home carrier's domain name, and a shared secret key, which is used for encryption. If the ISIM application is absent from the UICC, IMS AKA procedures can still be performed using the credentials contained on the UE's USIM.

A mobile-originated call is handled by an MMTel server. This is accomplished by setting the IMS communication service identifier (ICSI) in the communication invitation to point to the MMTel service, resulting in calls being looped by S-CSCF through the MMTel server in each subscriber's home network. In SIP protocol, which IMS relies on, call control and bearer (transport) control are essentially independent. In the context of the IR.92 profile, there are two situations in

[10]IMS MMTel service is a global standard based on the IMS, offering converged, fixed, and mobile real-time multimedia communication using the media capabilities, such as voice, real-time video, text, file transfer and sharing of pictures, audio and video clips. With MMTel, users can add and drop media during a session. One can start with chat, add voice, add another caller, add video, share media and transfer files, and drop any of these without losing, or having to end, the session. MMTel allows a single SIP session to control virtually all MMTel supplementary services and MMTel media. All available media components can easily be accessed or activated within the session. Employing a single session for all media parts means that no additional sessions need to be set up to activate video, to add new users, or to start transferring a file. Although it is possible to manage single-session user scenarios with several sessions, for example, using a circuit-switched voice service that is complemented with a packet-switched video session, a messaging service, or both, there are some benefits in taking MMTel's single-session approach [36].

Samsung Ex. 1005, Page 122 of 1147

which it is desirable to have some communication between the two domains. The first is during call session establishment. It is possible for a call-terminating device to accept a call invitation (and to alert the terminating subscriber). This is not normally a problem in legacy circuit-switched systems because call and bearer control are relatively more coupled. In order to minimize the possibility of this occurrence, VoLTE requires the use of SIP pre-conditions framework. This ensures that the terminating subscriber is not notified of the incoming call unless it is certain that there are sufficient resources to support the call (with the required QoS). The second situation in which call and bearer control require some communication is in the context of call drops and call establishment failures. If, for example, communication between the LTE/System Architecture Evolution (SAE) network and the Internet is lost, or if the radio link fails, IR.92 requires the IMS call session to be reestablished or cleared gracefully on the mobile device side and on the network side. VoLTE emergency calls must also be supported by the UE and the network according to regulatory requirements.

The IMS calls are processed by the subscriber's S-CSCF in the home network. The connection to the S-CSCF is, however, indirect and via P-CSCF. Since the P-CSCF varies with time and the mobile device's serving network, an important step in the enablement of voice calling capabilities is the discovery of the P-CSCF.

In legacy systems, telephony features are managed via dedicated signaling messages. Since call control in VoLTE uses completely different signaling, supplementary services and features in VoLTE are managed differently. IR.92 requires the use of XCAP protocol. Using XCAP, the UE connects to the MMTel server via HTTP and utilizes the standard PUT, GET, and DELETE methods to modify XML documents, which describe the supplementary services and their state.

3GPP adaptive multi-rate (AMR) and AMR-wideband (AMR-WB) voice codecs are specified for VoLTE. This choice simplifies interoperability with legacy 3GPP networks, in particular, in the context of transcoder-free operation. In order to ensure interoperability with all legacy telephony networks, in addition to the necessary voice transcoders, support for transmission of dual-tone multi-frequency (DTMF) signaling, which is used for telecommunication signaling over analog telephone lines, is mandatory.

VoLTE requires the use of a particular set of radio bearers, specifically, two signaling radio bearers (SRBs) and two or three data radio bearers (DRBs). The SRBs (SRB1 and SRB2) in E-UTRAN transport the dedicated control channel, which is used for RRC message exchange between the UE and the eNB, as well as NAS message exchange between the UE and the MME. These radio bearers always operate in RLC acknowledged mode. The two (or three) DRBs carry SIP (call control) and XCAP (service management) signaling as well as voice packets. The first DRB corresponds to the default EPS bearer and is used to transport SIP and XCAP signaling messages. The associated QCI index is 5, which means that the bearer operates in non-GBR mode but with high priority, low latency, and very low packet loss. The second DRB corresponds to a dedicated EPS bearer and is used to transport voice packets. The associated QCI index is 1, which

means that the bearer operates in GBR mode with high priority and low latency, but with relatively higher packet loss. There may also be an additional DRB operating with QCI index 8/9 (low-priority, high-latency, and very low packet loss) used for the transport of XCAP messages.

All networks and mobile devices are required to utilize a common APN for VoLTE, which in this case is IMS. Unlike the legacy networks, LTE networks employ an always-on concept which means that devices have PDN connectivity virtually from the moment they attempt their initial attach to the core network. During the initial attach procedure, some devices choose to identify the access point through which they prefer to connect. However, mobile devices are not permitted to name the VoLTE APN during initial attach, that is, to utilize the IMS as their main PDN, rather to establish a connection with the IMS application profile separately. Thus, VoLTE-enabled devices must support multiple simultaneous default EPS bearers. Note that because the VoLTE APN is universal, mobile devices will always connect through the visited PLMN's IMS P-GW.

We mentioned earlier that VoLTE sessions utilize two or three DRBs. This implies the use of one default EPS bearer plus one or two dedicated EPS bearers. The default EPS bearer is always used for SIP signaling, and exactly one dedicated EPS bearer is used for voice packets regardless of the number of active voice media streams. The XCAP signaling may be transported over its own dedicated EPS bearer or it may be multiplexed with the SIP signaling on the default EPS bearer, in which case only two EPS bearers are utilized.

The VoLTE-compatible mobile devices are required to operate in dual-stack IPv4/IPv6 mode (i.e., IPv4 and IPv6 networks run in parallel) by default [28]. If the IMS application profile assigns an IPv6 address, the device is required to prefer that address and specifically to utilize it during P-CSCF discovery. Since IP header overhead in VoIP calls can be considerable, VoLTE mandates IP header compression and the use of RoHC protocol for voice packets.

In the EPS network, mobility is implicitly handled for VoLTE by the defined mobility management mechanisms. A challenge arises during mobility to/from UMTS/GSM. UMTS/GSM utilizes a different call model than VoLTE. The call processing and service delivery entities are the MSC/Visited Location Register (VLR) and HLR rather than the CSCFs and MMTel servers. Furthermore, the bearer/transport channels used to exchange voice packets are circuit-switched and not packet-switched. Part of the solution to this problem is known as IMS centralized services. With IMS centralized services, voice calls are always anchored in the IMS, regardless of the domain from which they are originated. This solves the problem of transfer of call control as the mobile device moves between different radio access technologies; nevertheless, the packet-switched bearer challenge still remains. This problem inspired the concept of packet-switched over circuit-switched in UMTS and GSM, where a circuit-switched bearer is utilized to transport voice packets for calls anchored in the IMS. Voice call continuity during radio handover between LTE and the UMTS/GSM circuit-switched domain is managed by an SRVCC mechanism.

In conjunction with support for voice services, VoLTE-compatible devices also must support sending and receiving SMS over IP and instant messaging services. Furthermore, VoLTE-compatible networks must be capable of functioning as an IP short message gateway. Note that SMS over IP represents an alternative mechanism for providing short messaging services to users. Since mobile devices might support both of these SMS over LTE solutions, the operators are required to provision the preferred SMS transport via the IMS management object [53,64].

2.11.4 Circuit-switched fallback

Circuit-switched fallback (CSFB) is an alternative mechanism to move a subscriber from LTE to a legacy radio access technology in order to use circuit-switched voice services (see Figure 2.24). This function is only available, if LTE network coverage is overlapped by GSM, UMTS, or CDMA2000 coverage. The CSFB was already specified in 3GPP Rel-8, with further enhancements defined in later releases [5]. A number of different CSFB mechanisms are available and there are also differences depending on the radio access technology to which the subscriber falls back. The CSFB to GSM, UMTS, and CDMA2000 is defined, but no mechanisms are specified to support CSFB to both UMTS/GSM and CDMA2000 in the same PLMN. In this section, we describe the CSFB mechanisms for UMTS and GSM followed by CSFB procedure to CDMA2000. As mentioned earlier, depending on the target circuit-switched radio access technology, several CSFB solutions exist. As an example, there are three different solutions for CSFB to UMTS, two using the RRC connection release with redirection mechanism, and one using the packet-switched handover mechanism. For CSFB to GERAN, in addition to the RRC connection release with redirection and the packet-switched handover mechanisms, cell change order with or without network-assisted cell change can be used.

The CSFB affects the radio and the core networks. Figure 2.25 illustrates the modified EPS architecture with support for CSFB. It includes the interfaces between the different radio access network types and the core network entities. Note that UTRAN is the UMTS access network and GERAN is the GSM/EDGE access network. To support circuit-switched services, a connection to the MSC server must be established. The MME interfaces with the MSC server via the SGs interface. The CSFB mechanism is implemented using the SGs interface.

The legacy 2G/3G and the EPS can coexist in the form of hybrid networks in the transition period of network upgrades, where the MME serves LTE users, the SGSN provides service to UMTS users when utilizing data services, and the MSC server serves the users when utilizing circuit-switched voice services. The MSC server connects to the operator's telephony network. To support CSFB signaling and SMS transfer for LTE devices, the MME connects to the MSC server. The architecture shown in Figure 2.25 is a simplified schematic of the hybrid E-UTRAN and UTRAN/GERAN networks. The SGs interface between the MSC server and the MME enables the UE to register with both circuit-switched and

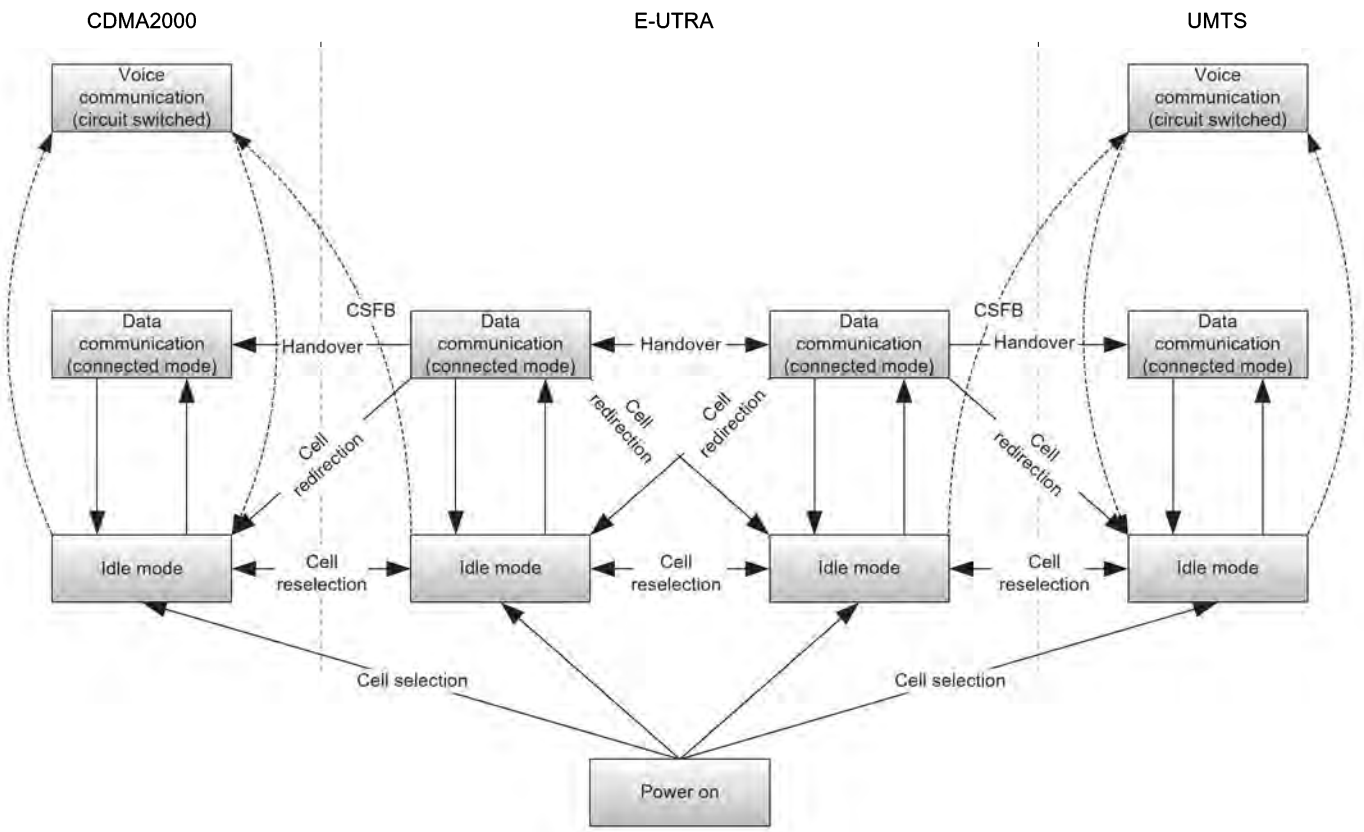


FIGURE 2.24

The concept of CSFB to GERAN or UTRAN [39].

packet-switched networks while on the LTE access network. This interface also enables the delivery of circuit-switched paging messages via the LTE access, as well as SMS, without having the device leave the LTE network. With the default LTE data network connection in operation, a mobile-terminated or incoming circuit-switched voice call triggers a paging message via the LTE network to the UE, as shown in Figure 2.26. The paging message initiates the CSFB operation, as the device sends an extended service request to the network to transition to UMTS/GSM access networks.

Upon transition to circuit-switched mode, the legacy call setup procedures are followed to set up the circuit-switched call. Mobile-originated or outgoing calls follow the same transition from packet-switched LTE to circuit-switched UMTS/GSM, except for the paging step, which is not needed. In UMTS networks, packet-switched data sessions can also move for simultaneous voice and data services. In GSM networks, packet-switched data sessions may be suspended until the voice call ends and the device returns to LTE mode, unless the GSM network supports dual transfer mode, which permits simultaneous voice and data. When the voice call ends, the device returns to LTE mode via idle mode or connected mode mobility procedures, as shown in Figure 2.26.

In EPS network architecture, the CSFB enables the provisioning of voice and traditional circuit-switched domain services. To provide these services, the EPS reuses circuit-switched infrastructure when the UE is served by E-UTRAN. A CSFB-enabled terminal, connected to E-UTRAN, may use GERAN or UTRAN to connect to the circuit-switched domain. This function is only available when the E-UTRAN coverage is overlapped by GERAN or UTRAN coverage. Figure 2.25 illustrates the architecture for CSFB in EPS. The CSFB and IMS-based services can coexist in the same operator's network. Although it is not very straightforward to support CSFB, all participating elements, i.e., UE, MME, MSC, and E-UTRAN must be modified to support additional functionalities.

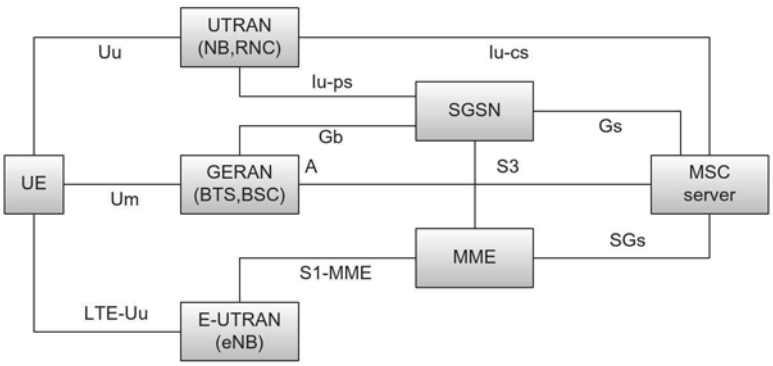


FIGURE 2.25

Modified EPS network architecture to support CSFB [5].

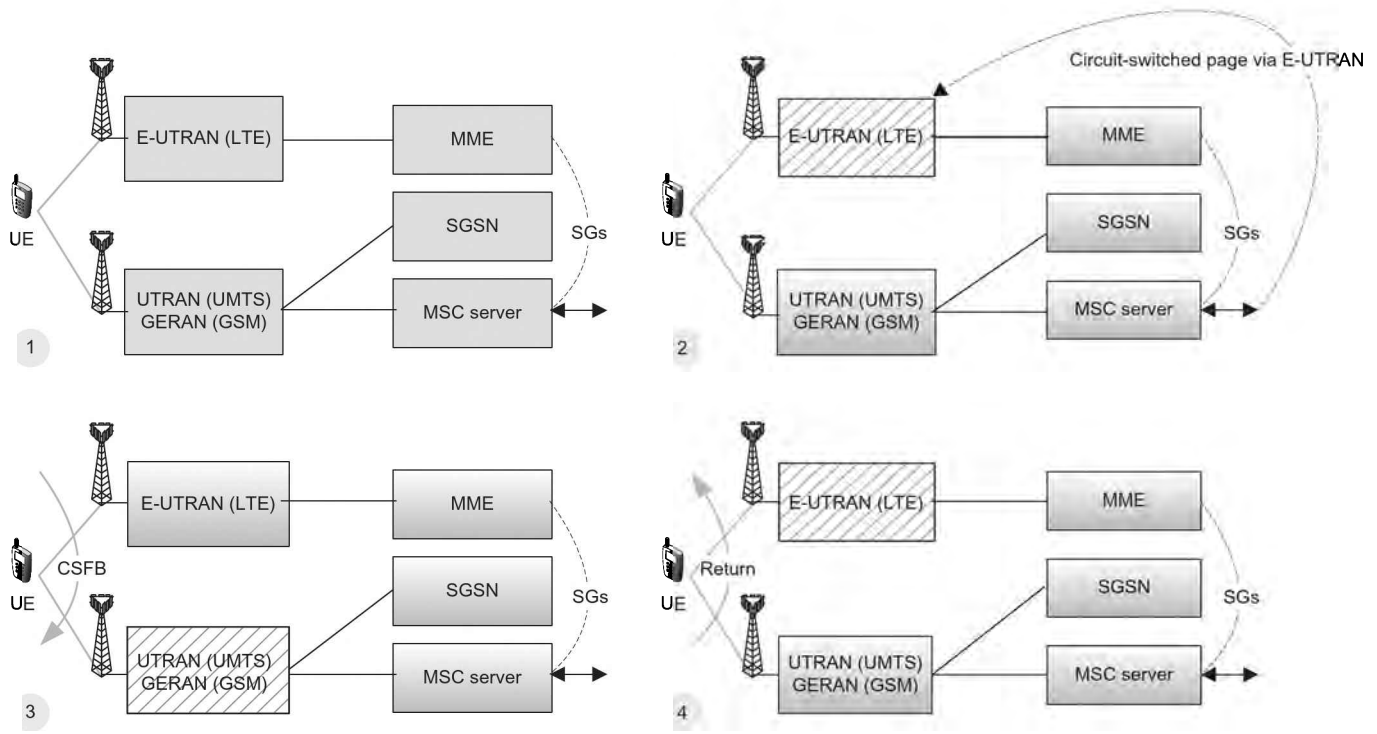


FIGURE 2.26

CSFB call processing in the hybrid network [63].

The CSFB support in EPS includes a new interface SGs, which is the reference point between the MME and MSC server. The SGs interface is used for mobility management and paging procedures between the EPS and the circuit-switched domain. The SGs reference point is also used for delivery of both mobile-originated and mobile-terminated SMS service. The CSFB-enabled network elements are required to support the following additional functions [54]:

- **UE:** Support access to E-UTRAN/EPC as well as access to the circuit-switched domain over GERAN and/or UTRAN; combined procedures for EPS/IMSI attach, update, and detach; and CSFB and SMS procedures for using circuit-switched domain services.
- **MME:** Deriving a VLR number and location area identity (LAI)[11] from the GUTI received from the UE or from a default LAI; maintaining of SGs association toward MSC/VLR for an EPS/IMSI attached UE; initiating IMSI detach at EPS detach; initiating paging procedure toward eNB when MSC pages the UE for circuit-switched services; support of SMS procedures; rejecting CSFB call request; and use of the LAI and a hash value from the IMSI to determine the VLR number when multiple MSC/VLRs serve the same LAI.
- **MSC:** Maintaining SGs association toward MME for an EPS/IMSI attached UE and support of SMS procedures specified by 3GPP.
- **E-UTRAN:** Forwarding paging request and SMS to the UE and directing the UE to the target circuit-switched-capable cell.

A CSFB and IMS-capable UE would follow the procedures for domain selection for UE originating session/calls. If a UE is configured to use SMS over IP services and it is registered to IMS, then it would send SMS over IMS, even if it is EPS/IMSI attached. The home operator has the option to activate/deactivate the UE configuration to use SMS over IP by means of device management in order to allow alignment with home PLMN support of SMS over IP. When UE performs CSFB procedure for a mobile-originated call for the purpose of an emergency call, it needs to indicate to the MME that this CSFB request is for an emergency purpose.

When the UE is paged via LTE with an incoming call, or when the user initiates an outgoing call, the device switches from LTE to UMTS/GSM mode. The acquisition of the UMTS/GSM network and setup of the call can be done in one of following two ways: handover or redirection. In the handover procedure, the target cell is prepared in advance and the device can enter that cell directly in the connected mode. Inter-RAT measurements of signal strength may be conducted by the UE while on the LTE system prior to executing the handover. In the redirection procedure, only the target frequency is indicated to the device.

[11]Each location area of a public land mobile network has its own unique identifier which is known as its LAI. This internationally unique identifier is used for location update of mobile subscribers. It is composed of a three decimal digit mobile country code, a two to three digit mobile network code that identifies the PLMN in that country, and a location area code which is a 16 bit number, thereby allowing 65,536 location areas within one PLMN.

The device is then allowed to select any cell on the indicated frequency, or may even try other frequencies/RATs, if no cell can be found on the target frequency. Once a cell is found, the device initiates normal call setup procedures. Inter-RAT measurements of signal strength are not needed prior to redirection. Consequently, CSFB with redirection may require less time to identify the best cell compared to the handover procedure [49,63,76].

Redirection-based CSFB has variations with differing call setup latencies. In 3GPP Rel-8 basic release and redirection approach, the UE follows LTE Rel-8 procedures and reads the system information which is broadcast through system information block (SIB) messages prior to accessing the target cell. The device can alternatively follow LTE Rel-8 procedures but may read only the mandatory SIBs 1, 3, 5, and 7, and skip other SIBs prior to access to the target cell. In the latter case, the neighbor cell information in SIB11 is delivered to the UE via measurement control messaging once the UE is in the connected mode on the target cell. This approach can be either implicitly supported by the UE and the network, or explicitly through deferred measurement report control signaling.

In 3GPP Rel-9 enhanced connection release with redirection and system information tunneling approach, the device follows LTE Rel-9 procedures, where SIB information can be tunneled from the target access network via the core network to the source access network, and can be included in the redirection message sent to the device. This can avoid reading any SIBs on the target cell. The solutions deployed today are predominantly based on 3GPP Rel-8 connection release with redirection with SIB skipping, in order to achieve acceptable call setup times, reliability, and to simplify deployments. While it is anticipated that 3GPP Rel-9 enhanced connection release with redirection will be deployed soon, a handover-based CSFB approach has not found any traction, since both release with redirection with SIB skipping and enhanced connection release with redirection with system information tunneling can address any call setup latency issues that may arise as a result of the basic release with redirection solution.

Another key issue for the voice call user experience is call setup reliability, which is the ability to successfully establish an incoming or outgoing call on the first attempt, or within a time frame that does not indicate call setup failure. The target is to at least match legacy performance, which is in the 98% range. With CSFB switching between LTE and UMTS networks, there are two primary challenges to call setup success: changes in inter-RAT conditions between measurement and acquisition for handover-based CSFB, and mismatches between LTE and 3G geographic signal coverage areas that require server-side updates. With handover-based CSFB, inter-RAT measurements can change between the time the measurement is conducted using LTE and the time UMTS/GSM/CDMA2000 voice network acquisition is attempted. During this time, the cell identified and prepared for handover may become unavailable, resulting in connection failure [63].

In handover-based CSFB, the measurement is conducted before the inter-RAT transition. If the inter-RAT conditions adversely change, there is a higher probability of handover failure, especially in high mobility scenarios. In practice, it is

shown the failure rate of inter-frequency handover or inter-RAT handover are relatively higher, suggesting that delays between inter-RAT measurement and network acquisition can increase setup failures. In contrast, redirection-based CSFB can result in higher call setup success rates relative to handover-based CSFB, because redirection-based CSFB solutions conduct inter-RAT measurements immediately before attempting access to the identified cell. Since the 3GPP Rel-9 system information tunneling and 3GPP Rel-8 redirection-based CSFB methods have shown satisfactory call setup times and higher reliability, the industry has taken the redirection-based CSFB approach [63].

CSFB to CDMA 1xRTT is supported for UEs with both single and dual receiver chain configurations. UEs with single receiver configuration are not able to camp in CDMA 1xRTT when they are active in E-UTRAN. The EPC provides mechanisms for the UE to perform registration procedure with CDMA 1xRTT, receive paging message, and send/receive SMS, while the UE is camped in E-UTRAN. UEs with dual receiver configuration can camp in CDMA 1xRTT while they are active in E-UTRAN, they may not be able to stay in E-UTRAN when they handle a circuit-switched call and/or perform registration signaling, and/or are sending or receiving SMS in CDMA 1xRTT.

The CSFB for CDMA 1xRTT in EPS enables the delivery of circuit-switched domain services (e.g., circuit-switched voice) by reusing CDMA 1xRTT circuit-switched infrastructure when the UE is served by the E-UTRAN. A CSFB-enabled terminal, while connected to E-UTRAN, may register in the CDMA 1xRTT circuit-switched domain in order to be able to use CDMA 1xRTT access to establish one or more circuit-switched services in the circuit-switched domain. The CSFB function is only available where E-UTRAN coverage overlaps with CDMA 1xRTT coverage. 3GPP specifications also specify the architecture for SMS over S102 interface in EPS. The mobile-originated or terminated SMS is tunneled in EPS over S102 and does not cause any CSFB to CDMA 1xRTT and consequently does not require any overlapped CDMA 1xRTT coverage. The CSFB to CDMA 1xRTT and IMS-based services can coexist in the same operator's network [5].

The CSFB to CDMA 1xRTT and SMS over S102 in EPS function are realized by reusing the S102 reference point between the MME and the 1xCS interworking solution (IWS) [16–18]. The reference architecture described in Figure 2.27 is similar to the SRVCC architecture for E-UTRAN to 3GPP2 1xCS, with the additional feature that the S102 session is extended similar to preregistration for S101. The S102 is the reference point between the MME and the 1xCS IWS which provides a tunnel between MME and 3GPP2 1xCS IWS to relay 3GPP2 1xCS signaling messages [14–18].

2.12 LIPA, selected IP traffic offload, and IP flow mobility

Femto-cells have been recently introduced to provide enhanced indoor coverage and to offload macro-cells. 3GPP Rel-10 network architecture introduced three

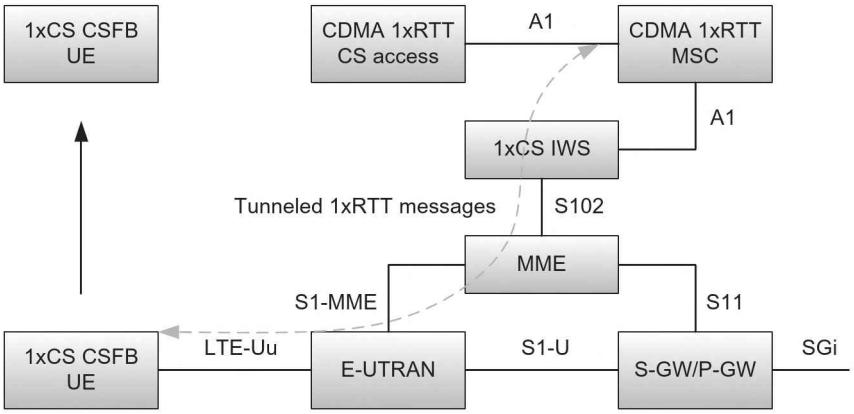


FIGURE 2.27

Reference architecture for CSFB to CDMA 1xRTT [5].

important concepts, namely, LIPA, selected IP traffic offload (SIPTO), and IP flow mobility [57,59].

2.12.1 Local IP access

LIPA is the ability for an IP-enabled UE to directly access a consumer's home local area network as well as the Internet (through the operator's IP core network) using the radio-link of a femto-cell or a home eNB (HeNB). The use of LIPA enables improved performance, more innovative services that interconnect mobile stations and home local area networks, and offloading of traffic from the operator's IP core network which is ultimately through the Internet. Enabling LIPA benefits mobile operators, subscribers, and Internet service providers. For mobile operators, the ability to offer LIPA via an HeNB would mean higher revenues by way of value-added services without having to invest significantly on upgrading network infrastructure to cope with the higher throughput and speed that such applications would demand. For subscribers, LIPA would mean faster and more secure data transfer, since traffic within the home network would not travel outside the subnet. Users could add customized high-speed applications such as file transfer, video streaming, and device sharing without involving the operator's core network, thus avoiding associated bottlenecks and possibly yielding lower data costs.

The requirements for LIPA can be generally classified into two main categories: LIPA to the home network and IP access to the Internet. For LIPA to the home network, the UE should be able to exchange data via the operator's core network and LIPA to the home-based network simultaneously, depending on the destination. The HeNB should decide, based on certain criteria, whether to route the data via the core network or to use LIPA. Furthermore, the UE should

be able to access other devices in the home network via the HeNB using the HeNB only as an access medium. Access to local IP through the HeNB E-UTRAN interface should only be granted to UEs with a valid subscription. This means that no specific modifications need to be done to the UE to support LIPA. In addition, it should be possible for a device in the home-based network to contact a UE via LIPA. Therefore, there is not any impact to the UE in accessing packet data services through the cellular core network, resulting in minimal changes to the protocol stacks in the core network.

IP access to the Internet should be possible without going through the operator's network. Simultaneous access from a UE to both the operator's core network and LIPA to the Internet are also possible. The operator or the HeNB owner, within the limits set by the operator, should be able to enable/disable LIPA to the Internet per HeNB. Furthermore, LIPA to the Internet should support a scenario where the HeNB and backhaul are provided by the same operator, as well as a scenario where the HeNB and backhaul are provided by different operators.

Basic functionality for LIPA was specified in 3GPP Rel-10. LIPA signifies the capability of a UE to obtain access to a local residential/enterprise IP network, subsequently called a local network, that is connected to one or more HeNBs. The LIPA functionality is being extended in 3GPP Rel-12 to allow access to the local network when a UE is in the coverage area of a macro-eNB and to provide mobility support (see Figure 2.28). LIPA allows a UE to work with devices in the local network, e.g., printers, video cameras, or a local web-server. LIPA allows the UE to discover supporting devices and to be discovered, and does not require the local network to be connected to the Internet but achieves IP connectivity with the UE through one or more HeNBs of the mobile operator. 3GPP Rel-10 only specified the support of LIPA when the UE accesses the local network via HeNB.

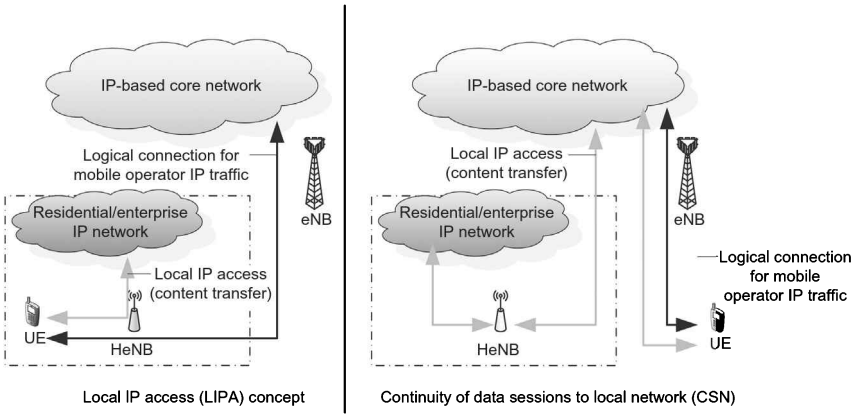


FIGURE 2.28

LIPA and continuity of data sessions concepts [73].

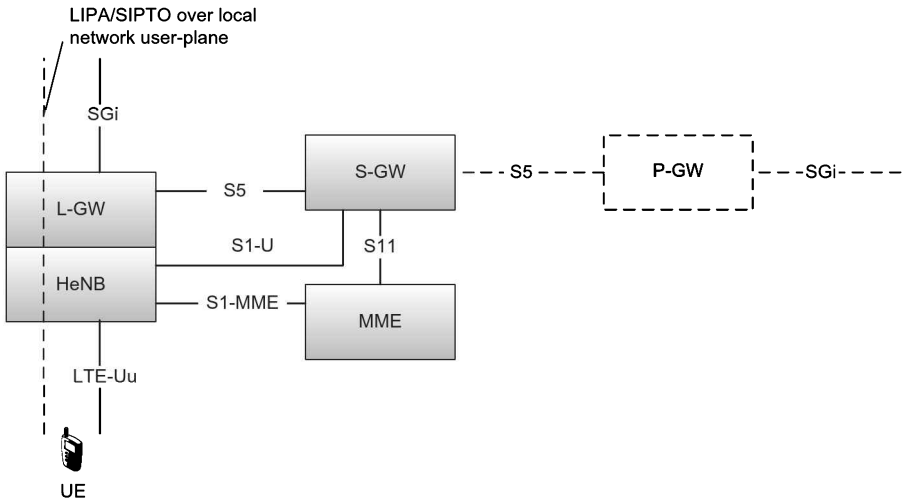


FIGURE 2.29

Network architecture to support LIPA or SIPTO over local network with L-GW collocated with the HeNB [73].

An operator may want to further provide access to the local network to a UE that is in the coverage area of a macro network. In 3GPP Rel-10, the UE was required to be able to maintain IP connectivity to the local network when moving between HeNBs within the same local network. However, access to the local network may be lost as a UE moves out of HeNB coverage into the macro network, even if other services (e.g., telephony, data services, and SIPTO) may continue after a handover to the macro network, resulting in an unsatisfactory user experience. A 3GPP Rel-12 study item will allow continuation of data sessions to the local network, when the UE moves between HeNB and the macro network (see Figure 2.28).

From an architectural perspective, a HeNB subsystem comprises a HeNB entity and it may optionally include a HeNB gateway (HeNB-GW) and a local gateway (L-GW). The LIPA and SIPTO functionalities over the local network are achieved using an L-GW that may be collocated with the HeNB. Figure 2.29 illustrates the network architecture to support LIPA and/or SIPTO over the local network with L-GW function collocated with the HeNB. It is shown that the HeNB subsystem is connected by means of an S1 interface to the EPC, and more specifically to the MME by means of an S1-MME interface and to the S-GW by means of an S1-U interface. When LIPA or SIPTO functions over the local network are activated (when the L-GW is collocated with the HeNB), the L-GW has an S5 interface with the S-GW. The local GW is the gateway toward the IP networks (e.g., residential/enterprise networks and Internet) associated with the HeNB. The L-GW performs some P-GW type functions including UE IP address allocation; DHCPv4 (server and client), and DHCPv6 (client and server)

Samsung Ex. 1005, Page 134 of 1147

functions; packet screening, as well as ECM-IDLE mode downlink packet buffering; and ECM-CONNECTED mode direct tunneling toward the HeNB [6].

The architecture for SIPTO over the local network with L-GW function collocated with a HeNB is depicted in Figure 2.29. For LIPA traffic, the addition of an L-GW function in EPS that can be either collocated with the HeNB or as a standalone node can provide partial P-GW and S-GW downlink data buffering functions. The S-GW in the core network serves the core network traffic. The L-S11 interface between the L-GW and the MME is used to manage the session for LIPA traffic. For the LIPA PDN connection, the L-GW needs to be located close to the HeNB, and to establish the connection through an L-S11 interface. When the UE is in the idle mode, the LIPA downlink packets are buffered in the L-GW. Upon UE transition to the connected mode, the packets buffered in L-GW are forwarded over the path between the L-GW and the HeNB. It is assumed that the user-plane does not go through the HeNB-GW for the direct tunnel [6].

The LIPA is realized using the L-GW entity that may be collocated with the HeNB. LIPA is established when the UE requests a new PDN connection to an APN for which LIPA functionality is permitted. In this case, the network selects the L-GW associated with the HeNB and enables a direct user-plane path between the L-GW and the HeNB. A HeNB supporting the LIPA function includes the L-GW address in the UE-related control messaging with the MME. The protocol option (i.e., GTP or PMIP) supported on the S5 interface between L-GW and S-GW is configured by the MME. As of 3GPP Rel-11, no interface between the L-GW and the PCRF is specified, and there is no support for dedicated bearers on the PDN connection used for LIPA. The L-GW rejects any UE requests for bearer resource modification.

The direct user-plane path between the HeNB and the collocated L-GW is enabled with a correlation ID parameter that is associated with the default EPS bearer on a PDN connection used for LIPA. Upon establishment of the default EPS bearer, the MME sets a correlation ID equal to the P-GW TEID (GTP-based S5) or the P-GW generic routing encapsulation (GRE) key (PMIP-based S5). The correlation ID is then signaled by the MME to the HeNB as part of E-RAB establishment and is stored in the E-RAB context in the HeNB. The correlation ID is used in the HeNB for matching the radio bearers with the direct user-plane path connections from the collocated L-GW. If the UE is roaming and if the HSS allows LIPA roaming for this UE in the visited PLMN, then the MME may provide LIPA for this UE.

LIPA is supported for APNs that are valid only when the UE is connected to a specific CSG. The MME releases a LIPA PDN connection to an APN, if it detects that the UE's LIPA CSG authorization data for this APN has changed and the LIPA PDN connection is no longer allowed in the current cell. As mobility of the LIPA PDN connection is not supported in 3GPP Rel-11, the LIPA PDN connection is released when the UE moves away from the HeNB coverage area. Prior to handover procedure, the HeNB requests through an intra-node signaling with the collocated L-GW to release the LIPA PDN connection. The HeNB determines

that the UE has a LIPA PDN connection from the presence of the correlation ID in the UE E-RAB context. The L-GW then initiates and completes the release of the LIPA PDN connection using the P-GW-initiated bearer deactivation procedure. During the handover, the source MME checks whether the LIPA PDN connection has been released. If it has not been released and the handover is an S1-based or an inter-RAT handover, the source MME rejects the handover. The direct signaling from the HeNB to the L-GW is only possible since mobility of the LIPA PDN connection is not supported. During idle state mobility events, the MME/SGSN deactivates the LIPA PDN connection, when it detects that the UE has moved away from the HeNB coverage area [6].

2.12.2 Selected IP traffic offload

The SIPTO refers to the ability to selectively forward different types of traffic via alternative routes to/from the terminal. Home network flavor of SIPTO depends on a HeNB. In this mode, specific traffic, determined by the operator's policy and/or subscription, is transferred to/from eNB directly to Internet/intranet, bypassing the mobile operator access/core network. Macro network flavor of SIPTO covers the ability to offload traffic next to an RNC or S-GW as opposed to traversing the operator's core network. SIPTO extends what LIPA aims to achieve by providing an alternate path for selected traffic. The primary objective is to add capacity to the primary RAN-core network path and reduce the unit cost of transport.

SIPTO for HeNB is similar to LIPA in terms of necessary changes to network capabilities and additional functionality in the form of an L-GW collocated with HeNB (Figure 2.30). Considering substantial complexity and limited benefit of SIPTO for HeNB as opposed to LIPA, 3GPP decided to mark that as a future study item beyond Rel-10. The SIPTO function enables an operator to offload certain types of traffic at a network node close to the UE's point of attachment to the access network. SIPTO beyond E-UTRAN can be achieved by selecting a set of network gateways (S-GW and P-GW) that is geographically/topologically close to a UE's point of attachment. SIPTO beyond E-UTRAN corresponds to a traffic offload through a P-GW located in the mobile operator's core network. SIPTO applies to the non-roaming, provided that appropriate roaming agreements are in place between the operators, and the roaming cases. Offload of traffic for a UE is available for UTRAN and E-UTRAN accesses only. When the UE enters to UTRAN/E-UTRAN from another type of access network (e.g., from GERAN), it is the responsibility of the new SGSN/MME to decide whether to perform deactivation with reactivation request for a given PDN connection, depending on whether SIPTO is permitted for that APN. Realization of SIPTO beyond E-UTRAN relies on the same architecture models and principles that govern the local breakout.

In order to select a set of appropriate network gateways (S-GW and P-GW) based on geographical/topological proximity to the UE, the gateway selection function uses the UE's current location information. In order to allow or prohibit SIPTO on per user and per APN basis by an operator, subscription data in the

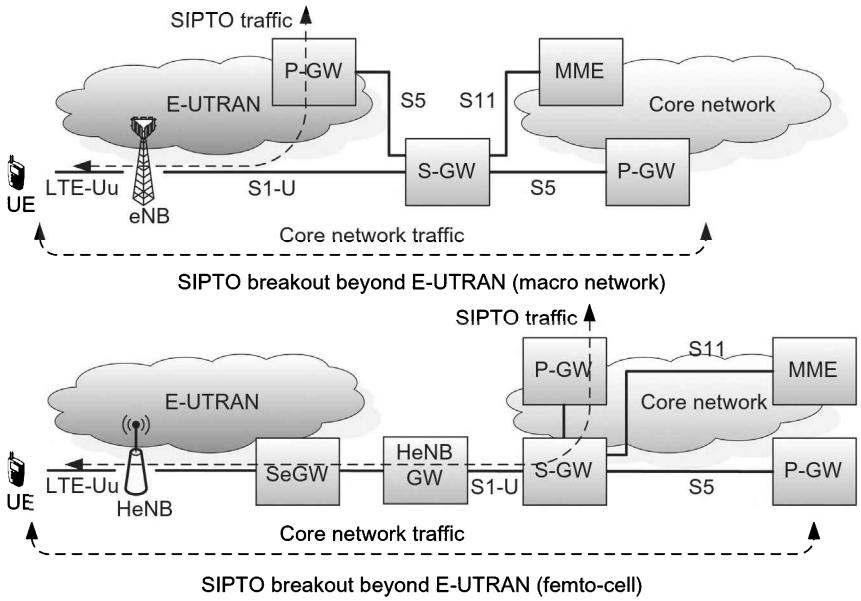


FIGURE 2.30

Illustration of SIPTO variants [62].

HSS is configured to indicate to the MME if offload is allowed or prohibited. If the SIPTO permission information from the HSS conflicts with MME's configuration for that UE, then SIPTO is not used. If HSS indicates the visited PLMN address is not allowed, then the MME does not provide SIPTO. The MME may be configured on an APN basis in order to use SIPTO (e.g., to handle the case where the HSS is not configured with SIPTO information for the UE). For SIPTO beyond E-UTRAN, as a result of UE mobility (e.g., detected by the MME at tracking area update procedure), the target MME may redirect a PDN connection toward a different gateway that is more appropriate for the UE's current location.

The SIPTO at the local network function enables an IP-capable UE connected via a HeNB to access a specific IP network or the Internet without the user-plane going through the mobile operator's network. The subscription data in the HSS is configured per user and per APN to indicate to the MME if offload at the local network is allowed. SIPTO at the local network can be achieved by selecting an L-GW function collocated with the HeNB or selecting stand-alone network gateways (with S-GW and L-GW collocated) residing in the local network. In both cases, the selected IP traffic is offloaded via the local network. Specific to the HeNB sub-system, the applicability of SIPTO at the local network does not depend on CSG membership and the feature can be applied to any UE, as long as the HeNB is configured for open or hybrid access mode. There is no interface between the L-GW and the PCRF, and there is no support for dedicated bearers

on the PDN connection used for SIPTO at the local network. The L-GW rejects any UE request for bearer resource modification. In 3GPP Rel-11, SIPTO at the local network is intended for offloading Internet traffic only, thus the L-GW does not provide APN-specific connectivity. Therefore, if the subscription data in the HSS indicates that offload at the local network is allowed, this implies that the related APN is typically used for providing Internet connectivity.

2.12.3 IP flow mobility

An IP flow consists of a set of packets that are exchanged between two nodes and that match a given flow description. In 3GPP Rel-8 mobility architecture, the granularity of access system connectivity was per PDN basis, that is, the UE could not access the same PDN via multiple interfaces, which meant that the UE could not treat individual IP flows within a PDN connection separately; therefore, while performing handover from one access network to another, it had to move all the IP flows within a PDN connection together [57]. In other words, in Rel-8 EPS, the UE could not communicate using multiple radio access networks simultaneously. While the UE could establish a single PDN connection or multiple simultaneous PDN connections, all traffic exchanged by the UE, regardless of the PDN connection it belonged to, was routed through the same radio access network.

IP flow mobility introduced the concept of treating IP flows individually within a PDN connection. It specified that the mobility of a PDN connection is handled per IP flow, which meant that within a PDN connection, the UE could establish IP flows over multiple access networks; selectively remove IP flows from an access system; selectively transfer IP flows between access systems; and transfer all IP flows from a certain access system. The following assumptions are valid while defining the functionality of multi-access PDN connectivity and IP flow mobility: (1) the number of non-3GPP interfaces simultaneously active with a 3GPP interface is limited to one; and (2) all IP flows are toward one, and only one PDN. However, in this framework a single IP flow cannot be split and routed via different access networks, and IP flows cannot be established toward multiple PDNs. The following use cases have been identified for IP flow mobility in 3GPP:

- Adding IP flow to an existing one via a different access network: The UE is reachable through both 3GPP and non-3GPP access networks. It has an active IP flow through the 3GPP (or non-3GPP) access networks. At some point in time, a new IP flow is initiated through non-3GPP (or 3GPP) access toward the same PDN. The non-3GPP (or 3GPP) access of the UE may or may not have any ongoing session through it, as depicted in Figure 2.31.
- Removing IP flow established via a different access network: The UE is reachable through both 3GPP and non-3GPP access networks and has simultaneously active sessions through both RANs. Each of the RANs can have one or more ongoing sessions. At some point in time, one IP flow via non-3GPP (or 3GPP) access is removed.

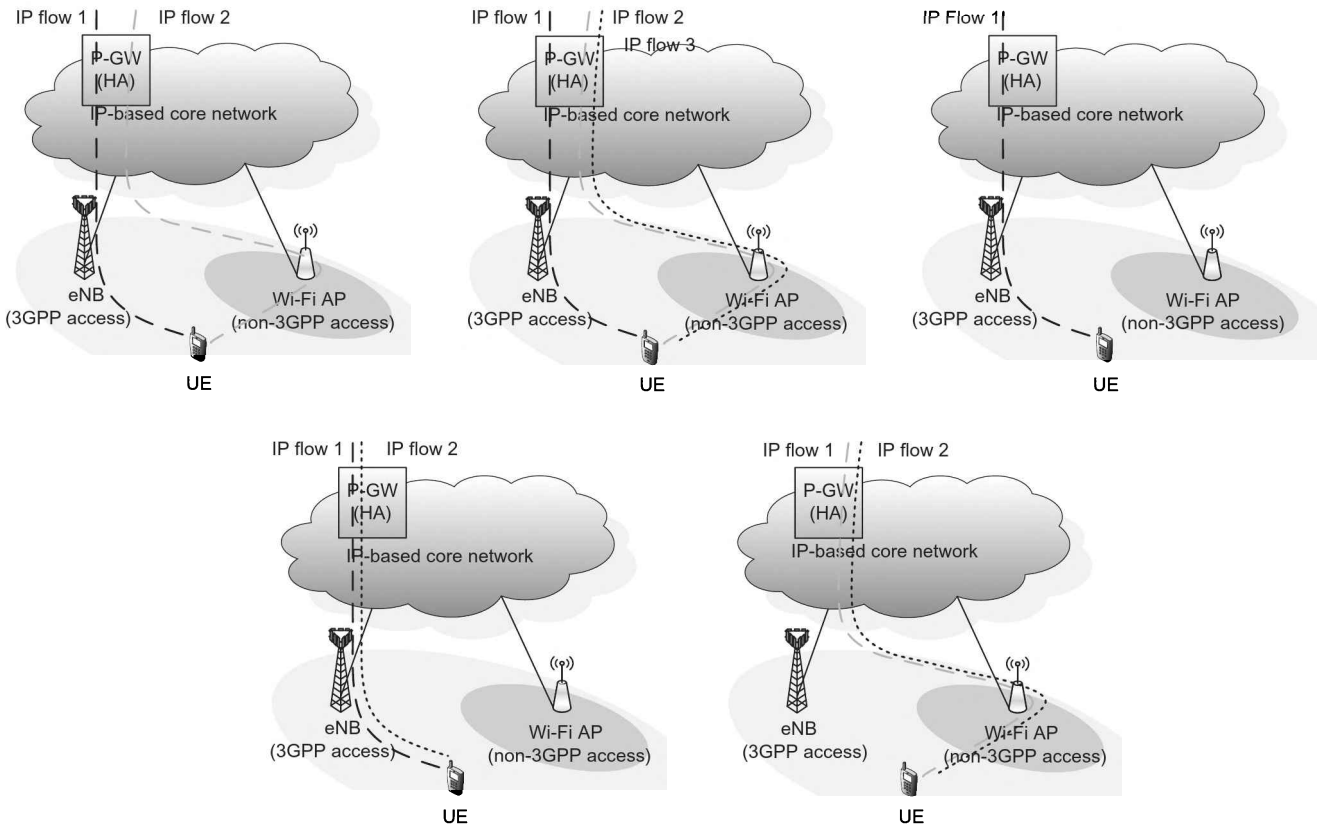


FIGURE 2.31

Adding, removing, and transferring of IP flows in various IP flow mobility scenarios using 3GPP and non-3GPP access networks [57].

- IP flow mobility between access networks: this use case involves selective transfer of IP flows between RANs. The state of the UE's interfaces can be one of the followings: (1) both radio interfaces are active simultaneously; or (2) only one radio interface is active. In the case of both radio interfaces being active simultaneously, the UE is reachable through both 3GPP and non-3GPP RANs and has simultaneously active sessions through both RANs. It has two IP flows through the 3GPP (or non-3GPP) access and one IP flow through the non-3GPP (or 3GPP) RAN. At some point in time, one IP flow is moved from the 3GPP (or non-3GPP) RAN to the non-3GPP (or 3GPP) RAN. The use case is depicted in Figure 2.31. In the case of only one active radio interface, the UE is camping on a 3GPP (or non-3GPP) RAN with at least two active IP flows. The UE is moving toward the coverage of a non-3GPP RAN. At some point in time, the UE detects a non-3GPP (or 3GPP) RAN and moves a subset of active IP flows (i.e., one IP flow) to the newly detected RAN.
- Moving all IP flows from one access network to another: The UE is reachable through both 3GPP and non-3GPP RANs. It has one or more active IP flows through the non-3GPP access. The 3GPP access of the UE may or may not have any ongoing session. The UE is moving away from the coverage of the non-3GPP RAN. At some point in time, due to loss of coverage of the non-3GPP RAN, the UE moves all IP flows going through the non-3GPP access to the 3GPP access.

As mentioned earlier, the granularity of access network connectivity and inter-system mobility is per PDN connection basis. This implies that when a handover occurs, all IP flows belonging to the same PDN connection are moved from the source access system to the target access system. With IP flow mobility, it is possible to have a finer granularity in access network connectivity and inter-system mobility. The handover procedures can be applied to a single or multiple IP flows belonging to the same PDN connection. This implies that some IP flows of one PDN connection can be routed via one access network, while simultaneously some IP flows of the same PDN connection can be routed via another access system. To achieve IP flow mobility, the inter-system mobility signaling is enhanced in order to carry routing filters. The extensions to DSMIPv6 mobility signaling are needed to carry routing filters when the UE is connected to multiple access networks simultaneously [4].

There are two possible approaches to IP flow mobility, client-based and network-based. The former relies on an IP host-centric solution, introducing a mobility client in the host and a mobility agent in the core network. The latter relocates the IP mobility client functionality from the host to the network, thus making the mobile device agnostic to any IP mobility signaling. Client-based IP mobility solutions require the user terminal to be involved in the management of the mobility, running a specialized stack that is able to detect, signal, and react to changes of point of attachment. Dual stack for mobile IPv6 is standardized by the IETF to provide client-based IP mobility support [29]. Network-based IP mobility solutions locate the mobility management control of the terminal in the network.

Samsung Ex. 1005, Page 140 of 1147

In this way, the terminal is not required to perform any kind of signaling (e.g., binding updates) to react to changes of its point of attachment to the network, as these changes are transparent to the mobile terminal IP stack. Proxy MIPv6 (PMIPv6) is the protocol standardized by the IETF to provide network-based IP mobility support [59].

2.12.4 Wi-Fi offloading

Wi-Fi offload offers a new way to extend the capacity and coverage of an LTE network. Wi-Fi networks are integrated to complement LTE networks and allow the use of each technology's advantages according to the actual demand. For example, the Wi-Fi is optimized for indoor environments, whereas the LTE network is designed for complete coverage in all areas. Depending on the trust relationship, the Wi-Fi may be integrated as a trusted or a non-trusted access technology. In the newest releases of 3GPP network architecture, a paradigm change has occurred. Instead of a handover for the complete connection, single traffic flows are routed over one access technology while the remaining ones are routed over the other. Therefore, the elaborated QoS feature in LTE may be used for delay or jitter sensitive IP flows, such as voice or video conference, whereas less time critical services may be routed over the cost-effective Wi-Fi when available. Corresponding handover procedures for the partial offload of certain flows are controlled by the network operator to ensure best user experience.

The user authentication is done using USIM credentials and according to LTE authentication procedures. Therefore, in the ideal case, the user does not perceive the offload and only recognizes an enhanced data rate using the mobile services. The Wi-Fi offloading refers to a scenario where part of the IP traffic to or from a device is routed via Wi-Fi access and another part via 3GPP access. Wi-Fi offloading covers both scenarios where the IP traffic via Wi-Fi radio is anchored in the EPC (i.e., seamless offloading) and the scenario where it is not anchored (i.e., non-seamless offloading). There is an access network discovery and selection function (ANDSF) entity to provide the UE with the access network discovery information and the policy on how to use the available access networks, available access networks, and preferred routing of the IP traffic per APN and per IP flow [58,61].

3GPP Rel-8 introduced a client-server-based protocol, dual stack mobile IP (DSMIP), to enable seamless handover between 3G and Wi-Fi. DSMIP is a mobility protocol specified in IETF that provides IP address preservation for IPv4 and IPv6 sessions, allowing the user to roam independently in IPv4 and IPv6 accesses. The solution does not require any support from Wi-Fi access network. If the cellular radio access network supports a home agent (HA), the only new requirement is that the UE (i.e., client) and the HA (i.e., server) are dual-stack capable. Applications and accesses can be either IPv4 or IPv6 compliant depending on the deployment models of the operators. Therefore, the changes required are only made to the UE and GGSN/P-GW. The HA can be implemented as a functionality in the GGSN/P-GW or a stand-alone entity. The HA is the anchor

that binds the permanent identifier of the node (i.e., home address or HoA) with the local address based on the node's location (i.e., care-of-address or CoA). The exposed IP address accessible by the application remains the same.

The DSMIP can be used over trusted and/or untrusted non-3GPP access (e.g., Wi-Fi) and/or 3GPP access (e.g., HSPA and LTE). Wi-Fi mobility in 3GPP Rel-8 can be implemented either using DSMIP over an H1 or S2c interface. The H1 interface is compatible with pre-Rel-8 GGSN and PDG, and the S2c interface is part of the EPS/LTE architecture. The DSMIP provides a better user experience compared to application-based switching, since it preserves the IP address whenever the network is changed. Furthermore, the DSMIP is easily portable to multiple devices since it does not depend on the application.

3G/Wi-Fi seamless offload is a further enhancement to the solution based on DSMIP that enables seamless handover between 3G and Wi-Fi. It also provides the possibility to move selected IP traffic (i.e., HTTP, video, and VoIP) while supporting simultaneous 3G and Wi-Fi access networks. The main components of the solution are DSMIP with extensions for seamless IP flow mobility, ANDSF extensions for providing operator control, and PCC enhancements for QoS [61]. The 3G/Wi-Fi seamless offload solution does not require any PCC support. When IP flow mobility is used and PCC function is deployed, the PCC architecture is enhanced to handle multiple simultaneous connections for a single IP session.

In dual radio scenarios, such as Wi-Fi and cellular, the two access networks are not aware of each other and one access network cannot control the device protocol state in the other access network. Also, Wi-Fi is not a controlled access network and works opportunistically in unlicensed spectrum; therefore, the only element in the network that is known is when a Wi-Fi is available, the radio quality and the performance that can be achieved through that Wi-Fi is available only to the device. As a result, the device is best positioned to decide the connectivity options. In this context, network discovery and selection procedures can be performed by a connection manager in the device. This kind of connectivity management is out of the scope of 3GPP and it is up to the vendors to provide a solution. More intelligent connection managers can search and prioritize the best available links based on predefined requirements, such as QoS, bandwidth, latency, jitter, power consumption, and operator policies. [61].

2.13 Mobile IP

Mobility support within EPS is based on a mobile IP framework. Mobile IP is an IETF protocol that allows mobile users to move from one network to another while maintaining their IP addresses [55]. The two versions of mobile IPs, i.e., Mobile IPv4 and Mobile IPv6, are described in IETF RFC 3344 [23] and IETF RFC 3775 [29], respectively. The mobile IP allows transparent routing of IP datagrams over the Internet. Each mobile node is identified by its home address

(HoA), irrespective of its current location in the Internet. When away from home, a mobile node is associated with a CoA, which provides information about its current location. The mobile IP specifies how a mobile node registers with its HA and how the HA routes datagrams to the mobile node through a tunnel. Using mobile IP, nodes may change their point-of-attachment to the Internet without changing their IP address, allowing the application and transport-layer protocols to seamlessly maintain connection while moving. The general characteristics of mobile IP can be summarized as follows:

- Transparency of user mobility to the transport and application layer protocols.
- Interoperability with stationary hosts running conventional IPs.
- Scalability across the Internet.
- Security by preventing an attacker from impersonating a mobile host.
- Macro mobility by ensuring long-term connection while away from HA.

Node mobility is realized without propagating host-specific routes throughout the Internet. Using mobile IP, a mobile device will have two addresses, that is, a primary or permanent HoA and a secondary or temporary CoA, which is associated with the network that the mobile node is visiting. There are two types of entities in mobile IP:

- An HA that stores information about mobile nodes whose permanent HoA is in the HA's network.
- A foreign agent (FA) that stores information about mobile nodes visiting its network. The FAs also advertise CoAs.

A node that wishes to communicate with the mobile node uses the permanent HoA of the mobile node as the destination address for outbound packets. Since the HoA logically corresponds to the network associated with the HA, conventional IP routing mechanisms forward these packets to the HA. Instead of forwarding these packets to a destination that is physically in the same network as the HA, the HA redirects these packets toward the FA. The HA looks for the CoA in the mobility binding table [55] and then tunnels the packets to the mobile node's CoA by appending a new IP header to the original IP packet, which preserves the original IP header. The packets are detected at the end of the tunnel by removing the IP header added by the HA and are delivered to the mobile node.

The mobile node directly sends packets to the other communicating node through the FA without involvement of the HA using its permanent HoA as the source address for the IP packets, that is, triangular routing. The FA can utilize reverse tunneling by sending the mobile node's packets to the HA, which forwards them to the communicating node. This mechanism is needed in networks whose gateway routers have ingress filtering enabled, and hence the source IP address of the mobile host needs to belong to the subnet of the foreign network; otherwise, the packets would be discarded by the router.

The mobile IP defines an authenticated registration procedure through which a mobile node informs its HA of its CoA, router discovery (which allows mobile

nodes to discover prospective HA and FAs), and the rules for routing packets to and from mobile nodes, including the specification of one mandatory and several optional tunneling mechanisms.

IP mobility enables routing data packets to the destination when moving to a foreign network. A central point is to maintain the IP address of a UE fixed so that there is no need for upper layers (above the IP layer) to adapt to the network change. Consequently, upper layer connections can continue without notification about the receiver's mobility. In general, IP mobility support can be realized with two different approaches: (1) client-based IP mobility; and (2) network-based IP mobility. This distinction does not depend on the underlying radio access technology; rather it is completely handled in the IP stack.

2.13.1 Client-based IP mobility

In the client-based mobility of IPv6, the UE carries the mobility extensions in its own IP stack. Central to this approach is to split the IP address into the HoA, which is the permanent IP address obtained from the home network, and the CoA, which corresponds to a temporary IP address and is obtained from the visited network. The administration of these addresses is done in an HA. As long as the UE is in the home network, the HA routes the data packets directly to the UE using the HoA. When the UE changes the network, it informs the HA with a proxy binding update (PBU) message about its new IP address, which is stored by the HA in the binding cache. This is essentially a lookup table which is searched for each incoming packet. If there is a CoA entry for this UE in the binding cache, the packet is instead forwarded to its CoA. As an optimization, it is also possible to route IP traffic between a correspondent node (CN) and the UE directly; in this case, a binding cache has to be additionally established in the CN itself.

In an extension, mixed networks with both IPv4 and IPv6 are supported with the dual stack mobile IPv6 protocol (DSMIPv6). This is essential to integrate most existing networks built on IPv4 into the new networks, which will be more and more equipped with IPv6 protocols. For a client-based mobility EPC access with Wi-Fi, DSMIPv6 is mandatory. With this approach, it is possible to offload data traffic from an LTE network to a Wi-Fi connection. However, this works only for a complete offload, that is, it is either possible to communicate over the LTE connection or over the Wi-Fi connection, but not over both. The reason is that in this architecture the Wi-Fi network is considered as a foreign network, to which all data packets are forwarded when there is a corresponding entry in the binding cache of the HA.

2.13.2 Network-based IP mobility

A completely different approach to support user mobility is the network-based IP mobility. Contrary to the client-based IP mobility, the network takes all necessary steps to route the data packets to the destination; thus, there is no need for the

client to do any signaling upon network change. There are two approaches for network-based IP mobility: PMIP version 6 (PMIPv6); and the GTP.

The PMIP (PMIPv6) is specified in IETF RFC 5213 [26]. Routing is based on two additional entities, the LMA, which works in a similar way as the HA in the client-based mobility approach, and the MAG, which implements the necessary mobility functions in the visited network. When the UE changes the network, the MAG is contacted by the new base station and informs the LMA about the change of location. Similar to the client-based mobility, this architecture has to be extended in order to implement the IP flow mobility capabilities, that is, moving selected IP flows from one access technology to another, and consequently, installing the required filters for flow routing. In order to send and receive data packets from and to any of its interfaces, the IETF has decided to adapt the logical interface (LIF). This is a software entity which hides the physical interface to the IP layer. This means that the mobile IP stack binds its sessions to the LIF, thus for the UE there is only one single interface to the IP network. The LIF controls the flow mobility in the UE. It is part of the connection manager in the operating system and has no impact on the IP stack. It represents a virtual interface which hides all flow mobility movements to the higher layers. A second aspect to introduce flow mobility to PMIPv6 is to provide signaling extensions to the MAG. This is necessary because the MAG will only forward traffic from and to a UE if the prefix has been delegated to the UE by this MAG. However, in IP flow mobility, this delegation might have been done by a different MAG before the flow handover. Signaling between the LMA and the target MAG resolves this issue.

3GPP has developed and specified the network-based mobility protocol called GTP. The PMIPv6 and client-based mobility were introduced as alternatives in LTE. The HA or LMA in the 3GPP EPC are located in the P-GW for both trusted and untrusted access. This is in contrast to the location of the MAG, where in the trusted access it is located in the Wi-Fi network and in the untrusted access is in the evolved packet data gateway (ePDG) network entity. Consequently, the connection between the LMA and the MAG can always be regarded as trusted, because the ePDG is under control of the EPC network operator. Up to 3GPP Rel-9 network architecture, the Wi-Fi offload was the only way to offload data from the EPC to Wi-Fi. The IP flow mobility was introduced with 3GPP Rel-10 network architecture for client-based mobility. Network-based mobility is not yet supported as of 3GPP Rel-11 network architecture and is being studied in 3GPP Rel-12.

2.14 Network architecture evolution and small cells

While macro-cellular networks will continue to provide essential wide-area coverage and support for high-mobility users, operators have started to look at other solutions to increase capacity in high-traffic areas. Wi-Fi offload is the most widely and successfully adopted solution in areas where subscriber density and

usage is high, such as urban areas and locations such as airports or stadiums. Operators are also exploring compatible solutions such as femto-cells to improve cellular coverage and capacity in residential areas, as well as small cell underlays to address the increasing demand for high capacity per area, and complement and strengthen their Wi-Fi and macro-cellular deployments. Mobile operators can further improve network utilization by actively managing the traffic beyond the radio access network within the core, using content caching, tiered pricing, and policy enforcement. These solutions do not increase the capacity per se, but make data transmission more efficient, allowing operators to process more content within the same network infrastructure. The integration of Wi-Fi and LTE small cells within the cellular core helps operators optimize network utilization across different radio access technologies (RATs), thus providing further improvement in performance, and creating a seamless multi-RAT experience for their subscribers [56,58,60,61]. While no single product or technology by itself can accommodate the current and future increase in data traffic, many viable solutions, working in concert, can be considered. The challenge for mobile operators is not the decision of which solution to select, rather it is how to best integrate multiple technologies within their networks, how to find the right balance to maximize their cumulative benefits, and how to leverage existing assets to facilitate the evolution of their networks.

Macro-cellular networks today typically provide coverage throughout a mobile operator's coverage area. Their capacity is adequate where the number of subscribers per cell site is relatively small. While high density areas account only for a small part of operators' footprint, they are of crucial importance. This is where subscribers are densely concentrated and where they use data services more extensively. The main challenge for mobile operators is to identify and correct capacity shortfalls in these high-traffic areas.

In order to increase network capacity and to meet the increasing demand for mobile wireless applications, a new approach is taken which relies on lower-power, shorter-range equipment, such as Wi-Fi access points or LTE/LTE-Advanced small cells installed closer to subscribers, in dense deployment areas. By limiting the range, the impact of interference is reduced and the capacity per area is increased. A single sector in a macro cell may have a comparable capacity to a Wi-Fi access point or a small cell, but it spreads it over a larger area, leading inevitably to lower capacity per area. Small cell deployments are one of the most viable strategies of mobile operators worldwide as they provide the capacity increase in the near term, along with the flexibility and compact form factor needed for highly localized deployments in high-traffic environments. Yet, there are several challenges that need to be addressed when planning for a small cell underlay network to enable the operator to maximize the benefits of the new network and to limit the costs. These include securing mounting assets where a high number of nodes are required and typically small cells are not mounted on telecommunication assets like cell towers that operators are accustomed to managing or leasing; backhaul for the small cells network for dense urban environments; as

well as integration with the cellular core network. The small cell underlay network must be fully integrated with the EPC and support self-organizing networks to give the operator the ability to manage user data traffic, support mobility, mitigate interference, and implement policy consistently across different radio access networks.

Most mobile operators no longer see Wi-Fi offload and small cells as competing solutions. They realize both are needed, with each playing an important role within a multi-RAT, multi-layer cellular network, spanning from macro base stations at one end, to residential femto cells at the other. Enhanced small cells can be deployed with macro coverage and stand-alone, both indoor and outdoor, and support both ideal and non-ideal backhauls. Enhanced small cells can also be deployed sparsely or densely. An enhanced small cell may benefit from the presence of overlaid macro cells, and it should also work without macro coverage, for example, in indoor scenarios. Cooperative mechanisms, between macro cell and small cell, as well as among small cells, may be beneficial for the above-mentioned use cases.

References

3GPP Specifications[12]

[1] 3GPP TS 23.002, Technical Specification Group Services and System Aspects; Network architecture (Release 11), December 2012.
[2] 3GPP TS 23.003, Numbering, Addressing and Identification (Release 11), December 2012.
[3] 3GPP TS 23.216, Single Radio Voice Call Continuity (SRVCC); Stage 2 (Release 11), December 2012.
[4] 3GPP TS 23.261, IP Flow Mobility and Seamless Wireless Local Area Network (WLAN) Offload; Stage 2 (Release 11), September 2012.
[5] 3GPP TS 23.272, Circuit-switched Fallback in Evolved Packet System; Stage 2 (Release 11), December 2012.
[6] 3GPP TS 23.401, General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (Release 11), December 2012.
[7] 3GPP TS 23.402, Architecture Enhancements for non-3GPP Accesses (Release 11), December 2012.
[8] 3GPP TR 23.882, 3GPP System Architecture Evolution: Report on Technical Options and Conclusions (Release 8), September 2008.
[9] 3GPP TS 29.274, Evolved General Packet Radio Service (GPRS) Tunneling Control Protocol for Control Plane (GTPv2-C) Stage 3 (Release 11), December 2012.
[10] 3GPP TS 24.301, Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3 (Release 11), December 2012.
[11] 3GPP TS 33.401, 3GPP System Architecture Evolution: Security Architecture (Release 11), December 2012.

[12]3GPP specifications can be found at the following URL: http://www.3gpp.org/ftp/Specs/archive/

[12] 3GPP TS 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 11), December 2012.
[13] 3GPP TS 36.413, Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S1 Application Protocol (S1AP), (Release 11), December 2012.

3GPP2 Specifications[13]

[14] 3GPP2 C.S0024-A v3.0, cdma2000 High Rate Packet Data Air Interface Specification, September 2006.
[15] 3GPP2 C.S0005-A v6.0, Upper Layer (Layer 3) Signaling Standard for CDMA2000 Spread Spectrum Systems, February 2002.
[16] 3GPP2 A.S0009-C v4.0, Interoperability Specification (IOS) for High Rate Packet Data (HRPD) Radio Access Network Interfaces with Session Control in the Packet Control Function, April 2011.
[17] 3GPP2 A.S0013-C v4.0, Interoperability Specification (IOS) for CDMA2000 Access Network Interfaces—part 3 Features, August 2012.
[18] 3GPP2 X.S0042-A v1.0, Voice Call Continuity between IMS and Circuit Switched System, August 2008.

IETF Specifications[14]

[19] IETF RFC 2460, Internet Protocol, Version 6 (IPv6) Specification, S. Deering, December 1998.
[20] IETF RFC 4960, Stream Control Transmission Protocol, R. Stewart, September 2007.
[21] IETF RFC 3310, Hypertext Transfer Protocol (HTTP) Digest Authentication Using Authentication and Key Agreement (AKA), A. Niemi, September 2002.
[22] IETF RFC 3261, SIP: Session Initiation Protocol, J. Rosenberg et al., June 2002.
[23] IETF RFC 3344, IP Mobility Support for IPv4, C. Perkins, August 2002.
[24] IETF RFC 2138, Remote Authentication Dial in User Service (RADIUS), C. Rigney et al., April 1997.
[25] IETF RFC 3588, Diameter Base Protocol, P. Calhoun et al., September 2003.
[26] IETF RFC 5213, Proxy Mobile IPv6, S. Gundavelli et al., August 2008.
[27] IETF RFC 2131, Dynamic Host Configuration Protocol, R. Droms, March 1997.
[28] IETF RFC 5555, Mobile IPv6 Support for Dual Stack Hosts and Routers, H. Soliman, June 2009.
[29] IETF RFC 3775, Mobility Support in IPv6, D. Johnson et al., June 2004.

Books

[30] S. Sesia, I. Toufik, M. Baker, LTE—The UMTS Long Term Evolution*: From Theory to Practice*, second ed., John Wiley & Sons, 2011.
[31] C. Johnson, Long Term Evolution in Bullets, second ed., Create Space Independent Publishing Platform, 2012.
[32] E. Dahlman, S. Parkvall, J. Skold, *4G, LTE/LTE-Advanced for Mobile Broadband*, Academic Press, 2011.

[13] 3GPP2 specifications can be found at the following URL: http://www.3gpp2.org/Public_html/specs/index.cfm

[14] IETF specifications can be found at the following URL: http://www.ietf.org/rfc.html

[33] T. Ali-Yahiya, Understanding LTE and Its Performance, Springer, 2011.

Articles

[34] LTE and Beyond: LTE, 4G, MME, P-GW, S-GW, Interfaces and beyond. <http://www.lteandbeyond.com/>.
[35] GPRS Tunneling Protocol, <http://en.wikipedia.org/wiki/GPRS_Tunnelling_Protocol>.
[36] (IMS) Multimedia Telephony (MMTel), from Wikipedia, <http://en.wikipedia.org/wiki/Multimedia_telephony>.
[37] K. Suzuki, et al., *Core network (EPC) for LTE*, NTT DoCoMo Tech. J. 13 (3) (2011).
[38] C.F. Lanzani, et al., Remote Radio Heads and the Evolution Towards 4G Networks, Altera, 2009<http://www.altera.com>.
[39] LTE System Description, <http://www.sharetechnote.com/>.
[40] LTE Identifiers, NMC Consulting Group. <http://www.nmcgroups.com>, 2011.
[41] LTE Identifiers, EventHelix.com Inc. <http://www.eventhelix.com/>, 2012.
[42] Small Cell Forum, <http://www.smallcellforum.org/resources-technical-papers>.
[43] G. Gómez, et al., *Towards aQoE-Driven resource control in LTE and LTE-A networks*, J. Comput. Netw. Commun, (2013).
[44] Common Public Radio Interface (CPRI), <http://www.altera.com/technology/high_-speed/protocols/cpri/pro-cpri.html>.
[45] GSMA, 2010, IR.92 IMS Profile for Voice and SMS V3.0. Available at: <http://www.gsma.com/newsroom/wp-content/uploads/2012/06/IR9230.pdf>.
[46] GSMA, 2011, IR.94 IMS Profile for Conversational Video Service V1.0. Available at: <http://www.gsma.com/newsroom/wp-content/uploads/2012/03/ir9410.pdf>.
[47] GSMA, 2011, IR.64 IMS Service Centralization and Continuity Guidelines V2.0. Available at: <http://www.gsma.com/newsroom/wp-content/uploads/2012/03/ir6420.pdf>.
[48] LteWorld, Evolution of Single Radio Voice Call Continuity. <http://lteworld.org/blog/evolution-single-radio-voice-call-continuity-srvcc>, December 2012.
[49] LteWorld, Understanding CS Fallback in LTE, September 2009.
[50] IP Multimedia Subsystem, from Wikipedia. <http://en.wikipedia.org/wiki/IP_Multimedia_Subsystem>.
[51] Network Switching Subsystem, from Wikipedia. <http://en.wikipedia.org/wiki/Network_Switching_Subsystem>.
[52] Internet Protocol Security (IPsec), <http://en.wikipedia.org/wiki/Ipsec>.
[53] C. Gessner, O. Gerlach, Voice and SMS in LTE, Rohde & Schwarz, 2011.
[54] Motorola White Paper, LTE Inter-technology Mobility: Enabling Mobility between LTE and other Access Technologies, Motorola Inc., November 2008.
[55] Mobile IP, from Wikipedia, <http://en.wikipedia.org/wiki/Mobile_ip>.
[56] T. Nakamura, et al., *Trends in small cell enhancements in LTE advanced*, IEEE Commun. Mag. (2013).
[57] T. Ahmed, et al., Multi access data network connectivity and IP flow mobility in evolved packet system (EPS), Proceedings of WCNC 2010, April 2010.
[58] A. Schumacher, J. Schlienz, WLAN Traffic Offload in LTE, Rohde & Schwarz, 2012.
[59] A. De La Oliva, et al., *IP flow mobility: smart traffic offload for future wireless networks*, IEEE Commun. Mag, 2011.

Samsung Ex. 1005, Page 149 of 1147

[60] Dealing with Density: The Move to Small-Cell Architectures, Ruckus Wireless White Paper, 2011.
[61] A 3G/LTE Wi-Fi Offload Framework: Connectivity Engine (CnE) to Manage Inter-System Radio Connections and Applications, Qualcomm White Paper, June 2011.
[62] R. Gupta, N. Rastogi, LTE-Advanced—LIPA and SIPTO, Aricent Group, 2012.
[63] Circuit-switched fallback, the first phase of voice evolution for mobile LTE devices, Qualcomm White Paper, 2012.
[64] Anritsu Whitepaper, Voice over LTE (VoLTE), Anritsu Corporation, September 2012.
[65] G. Bertran, The IP Multimedia Subsystem in Next Generation Networks, May 2007.
[66] N. Kottapalli, Diameter and LTE Evolved Packet System, Radisys White Paper, September 2011.
[67] D. Soldani, Bridging QoE and QoS for Mobile Broadband Networks, Huawei Technologies, 2010.
[68] M. Grega, et al., Quality of Experience Evaluation for Multimedia, Overview of Telecommunications, Telecommunication News, No. 4/2008, April 2008.
[69] S. Tripathi, et al., LTE E-UTRAN and its Access Side Protocols, Radisys White Paper, September 2011.
[70] Introduction to Evolved Packet Core, Alcatel-Lucent White Paper, 2009.
[71] The LTE Network Architecture: A comprehensive tutorial, Alcatel-Lucent White Paper, 2009.
[72] VoLTE with SRVCC: the second phase of voice evolution for mobile LTE devices, Qualcomm White Paper, October 2012.
[73] S. Kundalkar, N. Madhur Raj, LIPA: Local IP Access via Home Node B, Radisys White Paper, September 2011.
[74] Dynamic Host Configuration Protocol, from Wikipedia, <http://en.wikipedia.org/wiki/Dynamic_Host_Configuration_Protocol>.
[75] Open Base Station Architecture Initiative, from Wikipedia, <http://en.wikipedia.org/wiki/OBSAI>.
[76] Internet Key Exchange, from Wikipedia, <http://en.wikipedia.org/wiki/Internet_Key_Exchange>.

This page intentionally left blank

CHAPTER

E-UTRAN and EPC Protocol Structure

3

CHAPTER CONTENTS

3.1 EPS reference model and protocol structure 124
3.1.1 Functional split between E-UTRAN and EPC 127
3.1.2 Protocol stack for S1 user-plane 129
3.1.3 Protocol stack for S1 control-plane 129
3.1.4 Protocol stack for S5/S8 interface 131
3.1.5 Protocol stack for S11 interface 131
3.1.6 Protocol stack for X2 user-plane 132
3.1.7 Protocol stack for X2 control-plane 134
3.2 E-UTRAN and NAS protocols 135
3.3 LTE-Uu interface protocols 138
3.4 Support for relay nodes 144
3.5 Support for HeNB nodes 148
References 151

This chapter presents an overview of the LTE/LTE-Advanced protocol structure with the intent of describing where important functions reside in the E-UTRAN and EPC nodes, protocol layer terminations at various access and core network nodes, as well as the functional split between the E-UTRAN and the EPC. The E-UTRAN radio protocols can be separated into control-plane and user-plane entities, where the user-plane protocols are typically responsible for carrying user data, and control-plane protocols are used to transfer signaling and control information.

In general, a communications protocol is a system of rules for message exchange within or between network nodes. Individual protocols within a protocol suite are often designed with a single purpose in mind. This modularity facilitates the design and evaluation of protocols. Because each protocol module usually (virtually or physically) communicates with its peer entity across a communication link, they are commonly seen as layers in a stack of protocols, where the lowest protocol layer always deals with physical interaction of the hardware

LTE-Advanced.

© 2014 Elsevier Inc. All rights reserved.

Samsung Ex. 1005, Page 152 of 1147

across the communication link. Every higher layer protocol adds more features. User applications usually deal only with the top-most layers. In a practical implementation, protocol stacks are often divided into three major groups: media, transport, and applications.

In the context of protocol structure, we will frequently use the terms "service" and "protocol." It must be noted that services and protocols are distinct concepts. A service is a set of primitives or operations that a layer provides to the layer(s) with which it is interfaced. The service defines what operations a layer performs without specifying how the operations are implemented. It is further related to the interface between two adjacent layers. A protocol, in contrast, is a set of rules presiding over the format and interpretation of the information/messages that are exchanged by peer entities within a layer. The entities use protocols to implement their service definitions. Thus, a protocol is related to the implementation of a service. The protocols and functional elements defined by 3GPP standards correspond to all layers of the OSI[1] seven-layer network reference model. Figure 3.1 illustrates the mapping of the lower layer OSI protocols to those of the LTE/LTE-Advanced systems. As shown in the figure, what 3GPP considers as layer 2 and layer 3 protocols is mapped to the OSI data link layer. The higher layer protocols in the 3GPP stack are the application and transport layers. The presentation and session layers are often abstracted in practice.

According to ITU-T Recommendation X.200, there are seven layers, labeled 1 to 7, with layer 1 at the bottom of the protocol stack [18]. Each layer is generically known as an N layer. An N + 1 entity (at layer N + 1) requests services from an N entity at layer N. At each level, two entities (N-entity peers) interact by means of the N protocol by transmitting protocol data units (PDU). A service data unit (SDU) is a specific unit of data that has been passed down from an OSI layer to a lower layer, and which the lower layer has not yet encapsulated into a PDU. An SDU is a set of data that is sent by a user of the services of a given layer, and is transmitted semantically unchanged to a peer service user.

[1]Open systems interconnection (OSI) is a standard description or a "reference model" for the computer networks which describes how messages should be transmitted between any two nodes in the network. Its original purpose was to guide product implementations to ensure consistency and interoperability between products from different vendors. This reference model defines seven layers of functions that take place at each end of a communication link. Although OSI is not always strictly adhered to in terms of grouping the related functions together in a well-defined layer, many if not most products involved in telecommunication make an attempt to describe themselves in relation to the OSI model. OSI was officially adopted as an international standard by the International Organization of Standards (ISO) and it is presently known as Recommendation X.200 from ITU-T. The layers of OSI model are classified into two groups. The upper four layers are used whenever a message passes from or to a user. The lower three layers (up to the network layer) are used when any message passes through the host computer [17–19].

Samsung Ex. 1005, Page 153 of 1147

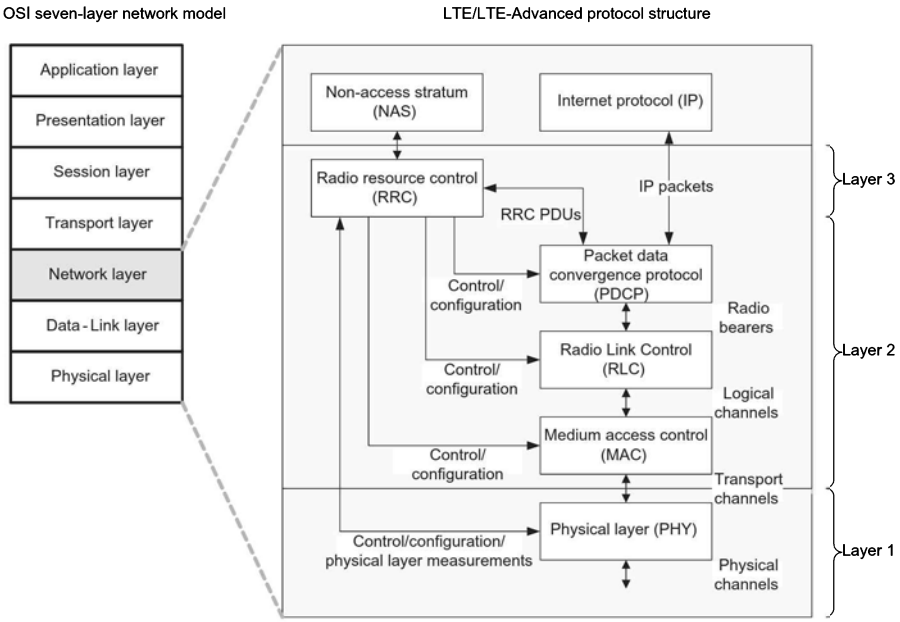


FIGURE 3.1

Mapping of lower E-UTRA protocol layers to OSI seven-layer network model.

The PDU at a layer N is the SDU of layer N − 1. In fact, the SDU is the payload of a given PDU. That is, the process of changing an SDU to a PDU consists of an encapsulation process, performed by the lower layer. All data contained in the SDU becomes encapsulated within the PDU. The layer N − 1 adds headers/sub-headers and padding bits (if necessary to adjust the size) to the SDU, transforming it into the PDU of layer N. The added headers/sub-headers and padding bits are part of the process used to make it possible to get data from a source to a destination.

As shown in Figure 3.2, a PDU is an information exchange between peer entities of the same protocol layer located at the source and destination. On the downward direction, the PDU is the data unit generated for the next lower layer. On the upward direction, it is the data unit received from the previous lower layer. An SDU, on the other hand, is a data unit exchanged between two adjacent protocol layers. On the downward direction, the SDU is the data unit received from the previous higher layer. On the upward direction, it is the data unit sent to the next higher layer.

This chapter provides a top-down systematic description of 3GPP EPC and E-UTRAN reference models, protocol structures, and functional decomposition, starting at the most general level and working toward details or specifics of the protocol layers, their functional constituents and interconnections. The main focus in this chapter is on the protocol aspects of LTE/LTE-Advanced. The functional details and implementation of the protocols will be described in Chapters 4–10.

Samsung Ex. 1005, Page 154 of 1147

3.1 EPS reference model and protocol structure

The evolved UMTS terrestrial radio access network consists of eNBs or equivalently E-UTRA base stations, providing the E-UTRA user-plane and control-plane protocol termination toward the UE, or alternatively, mobile station. As part of the evolution from and enhancement of the legacy UMTS systems, some radio network controller (RNC) functions have been included in the eNB to reduce the architectural complexity and further reduce the latency across the network. Figure 3.3 illustrates the EPS reference model comprising the network entities and the standardized interfaces between network elements. At a high level, the EPS encompasses EPC which is the core network and E-UTRAN, which is the radio access network. While the core network consists of many logical nodes, the access network is only made up a collection of eNBs, which are connected to the UEs via the LTE-Uu air-interface. Each of these network elements is

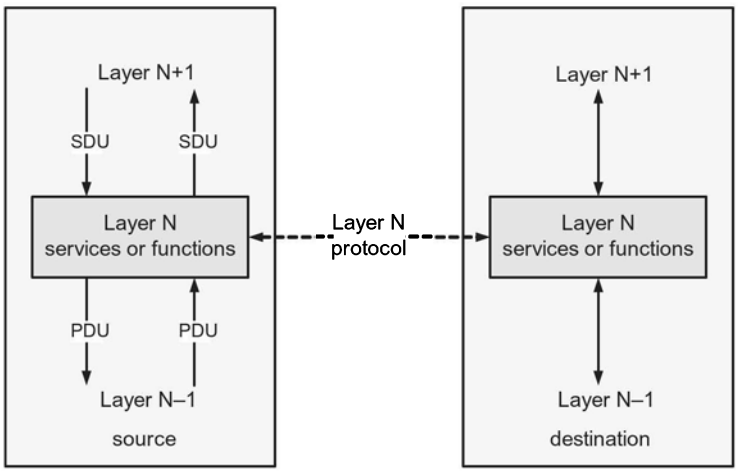


FIGURE 3.2

Illustration of service, protocol, PDU, and SDU concepts.

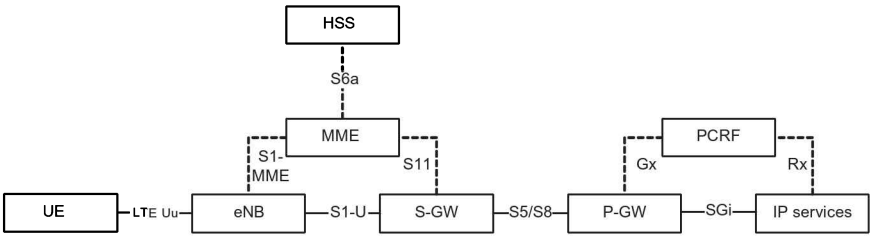


FIGURE 3.3

Reference model for EPS network elements and interfaces [9,18].

interconnected by means of interfaces that are standardized in order to facilitate interoperability among network equipment manufactured by different vendors.

In practice, a network operator, in its actual implementation/deployment, may split or merge some of the logical network elements depending on technical/economical/commercial considerations, but the overall functions remain the same. There are different options to split/combine MME, S-GW, and P-GW functionalities. Each option will result in different handling of control and user-plane packet routing. The choice may depend on operator's strategy and/or EPC vendor equipment availability and constraints. Integrating more logical entities into one physical entity will reduce latency, cost (fewer nodes and interfaces), and signaling overhead, but will increase the functional complexity per entity/component and will limit the benefit and potential savings achieved by a modular architecture, i.e., more redundancy, flexibility, and scalability. Hypothetically, a network implementation may begin with some level of integration of functional entities for simplicity, and then may be upgraded to more flexible and scalable architecture when EPS traffic grows.

In the reference model shown in Figure 3.3, the eNBs are typically responsible for radio resource management (RRM) that includes radio bearer control, radio admission control, connection management, dynamic allocation of resources to UEs in both uplink and downlink (i.e., scheduling); header compression and encryption of user payloads; selection of an MME at UE attachment when no routing to an MME can be determined from the information provided by the UE; routing of user-plane data toward S-GW; scheduling and transmission of paging messages (originated from the MME); scheduling and transmission of broadcast information (originated from the MME); and measurement and reporting for support of mobility and scheduling.

The MME is the key control-node for the 3GPP LTE access-network. It is responsible for idle-mode UE tracking and paging procedure including retransmissions. The MME functions include non-access stratum (NAS) signaling and security, inter-core-network signaling for mobility between 3GPP access networks, idle mode UE accessibility (including control and execution of paging retransmission), tracking area list management (for UE in idle and active mode), P-GW and S-GW selection, MME selection for handovers with MME change, roaming, authentication of the users, and bearer management functions including dedicated bearer establishment [1]. The NAS signaling terminates at the MME and it is also responsible for generation and allocation of temporary identities to the UEs. It verifies the authorization of the UE to camp on the service provider's PLMN and enforces UE roaming restrictions.

The S-GW routes and forwards user data packets, while also acting as the mobility anchor for the user-plane during inter-eNB handovers and as the anchor for mobility between LTE and other 3GPP technologies. The S-GW functions include local mobility anchor point for inter-eNB handover, mobility anchoring for inter-3GPP mobility, E-UTRAN idle mode downlink packet buffering and initiation of network triggered service request procedure, lawful interception, packet

routing and forwarding, and transport-level packet marking in the uplink and the downlink [1].

The P-GW provides connectivity of the UE to external packet data networks by acting as the point of exit and entry of traffic for the UE. A UE may have simultaneous connectivity with more than one P-GW for accessing multiple packet data networks. The P-GW functions include per-user packet filtering, lawful interception, IP address allocation to the UE, and transport-level packet marking in the downlink [1].

The policy control and charging rules function (PCRF) is responsible for enforcing network policy and flow-based charging functions in conjunction with the P-GW. The PCRF provides the QoS rules and provisions in accordance with the user's subscription profile to the policy and charging enforcement function (PCEF), which resides in the P-GW. The P-GW uses this QoS policy to assign the bearer-level QoS parameters. Note that in a typical scenario, multiple applications may be running in a UE at any time, each one having different QoS requirements. For example, a UE can be engaged in a VoIP call while at the same time browsing a web page or downloading an FTP file. VoIP has more stringent requirements for QoS in terms of delay and jitter compared to web browsing and FTP, while the latter requires much lower packet-loss rate. In order to support multiple QoS requirements, different bearers are set up within the EPS, each associated with a specific set of QoS attributes [8].

The home subscriber server (HSS) contains users' subscription data such as the EPS-subscriber QoS profile and any access restrictions for roaming. It also holds information about the PDNs to which the user can connect.

The reference model shown in Figure 3.3 also includes a number of standard reference points as follows [10]:

- S1-MME is a reference point for the control-plane protocol between E-UTRAN and MME.
- S1-U is a reference point between E-UTRAN and S-GW which is used for per-bearer user-plane tunneling and inter-eNB path switching during handover.
- S5 interface provides user-plane tunneling and tunnel management between the S-GW and P-GW. It is used for the S-GW relocation due to UE mobility and when the S-GW needs to connect to a non-collocated P-GW for PDN connectivity.
- S6a reference point enables transfer of subscription and authentication data for authenticating/authorizing user access between MME and HSS.
- Gx interface provides transfer of QoS policy and charging rules from PCRF to PCEF in the P-GW.
- S8 is an inter-PLMN reference point providing user- and control-plane between the S-GW in the visited PLMN and the P-GW in the home PLMN. The S8 interface is the inter-PLMN variant of S5.
- S11 is the reference point between MME and S-GW.
- SGi is the reference point between the P-GW and the packet data network.
- Rx is the reference point that resides between the application function and the PCRF.

The EPC and E-UTRAN network elements and their associated protocol layers are described in more detail in the following sections.

3.1.1 Functional split between E-UTRAN and EPC

The functional split between the EPC and E-UTRAN is shown in Figure 3.4. The main functions in each entity and their termination points in the network have been shown in the figure. An IP packet for a UE is encapsulated in an EPC-specific protocol and tunneled between the P-GW and the eNB for transmission to the UE. Different tunneling protocols are used across different interfaces. A 3GPP-specific tunneling protocol called GPRS tunneling protocol (GTP) is used over S1 and S5/S8 interfaces [7]. The protocol stack for the control plane between the UE and MME is shown in Figure 3.5. This includes both access stratum (AS) and NAS protocols. The lower layers perform the same functions as for the user-plane with the exception that there is no header compression function for the control-plane. The radio resource control (RRC) protocol is known as layer 3 in the AS protocol stack and in 3GPP RAN terminology [2]. It is the main controlling function in the AS responsible for establishing the radio bearers and

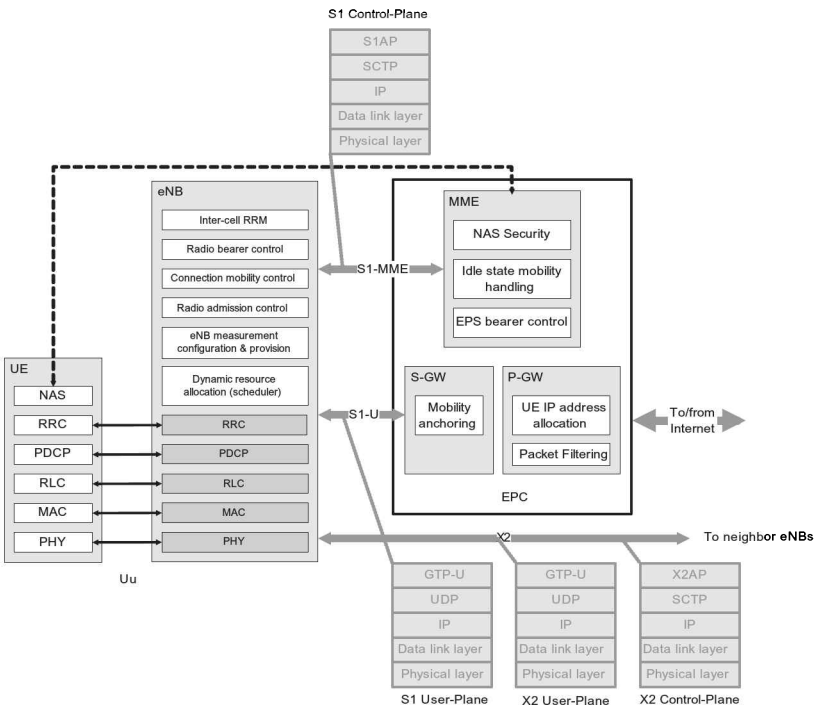


FIGURE 3.4

Protocol layers and functional split in UE, eNB, MME, S-GW, and P-GW [1].

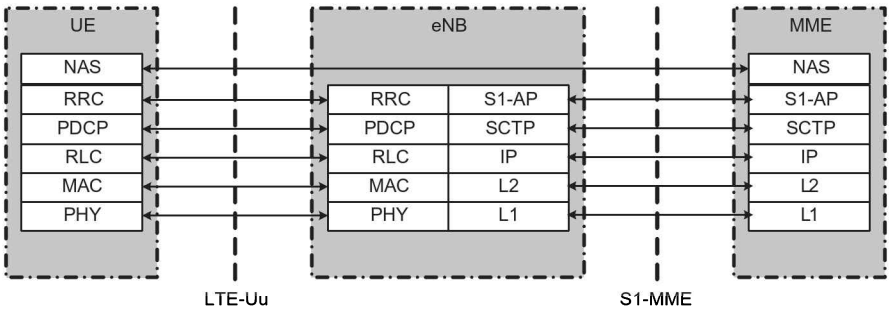


FIGURE 3.5

E-UTRAN control-plane protocol stack [10].

configuring all lower layers using RRC signaling between the eNB and the UE [16]. The E-UTRAN user-plane protocol stack is shown in Figure 3.6, consisting of the packet data convergence protocol (PDCP) [3], radio link control (RLC) [5], and medium access control (MAC) [4] sublayers that are terminated in the eNB on the network side. In the absence of any centralized control node, data buffering during handover due to user mobility in the E-UTRAN must be performed in the eNB itself. Data protection during handover is the responsibility of the PDCP sublayer. The RLC and MAC sublayers both restart in a new cell after the handover.

The S1 interface supports a multi-point connection among MMEs/S-GWs and eNBs. The S1-MME carries S1 application protocol (S1-AP) messages, using SCTP[2] over IP to provide guaranteed data delivery. Each SCTP association between an eNB and an MME can support multiple UEs. The S1-MME is responsible for EPC bearer setup and release procedures, handover signaling, paging, and NAS signaling transport. The S1-U consists of GTP-U protocol running on top of user datagram protocol (UDP), which provides best-effort data delivery. One GTP tunnel is established for each radio bearer in order to carry user traffic between the eNB and the selected S-GW.

[2]Stream control transmission protocol (SCTP) is a protocol for transmitting multiple streams of data at the same time between two end points that have established a connection in a network. SCTP is designed to facilitate a telephone connection over the Internet and specifically to support the telephone system's Signaling System 7—SS7—on an Internet connection. A telephone connection requires that signaling information, which controls the connection, be sent along with voice and other data at the same time. SCTP is also intended to make it easier to manage connections over a wireless network and to manage the transmission of multimedia data. SCTP is a standard protocol (IETF RFC 4960) developed by the IETF. SCTP manages reliable transport, ensuring complete arrival of data units that are sent over the network [13]. Unlike TCP, SCTP ensures complete and concurrent transmission of several streams of data (in units called messages) between connected end points. SCTP also supports multi-homing, which means that a connected end point can have alternate IP addresses [10,14].

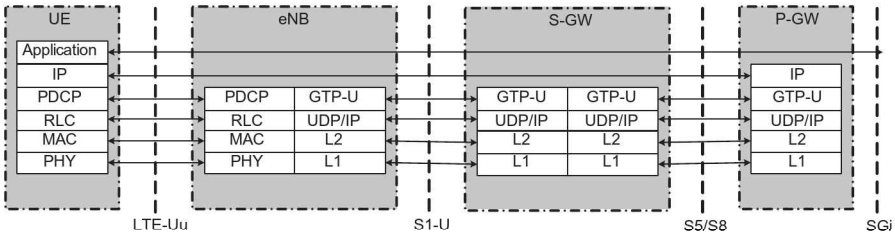


FIGURE 3.6

E-UTRAN user-plane protocol stack [10].

The GTP is a group of IP-based transport protocols that are used to carry IP packets within GSM, UMTS, and LTE networks [7]. As we discussed in Chapter 2, in 3GPP architecture, GTP and Proxy Mobile IPv6-based interfaces are specified on various interface points [12]. GTP can be decomposed into separate protocols: GTP-C and GTP-U. GTP-C is used within the GPRS core network for signaling between gateway GPRS support nodes (GGSN) and serving GPRS support nodes (SGSN). This allows the SGSN to activate a session on a user's behalf (PDP context activation), to deactivate the same session, to adjust QoS parameters, or to update a session for a subscriber who has just arrived from another SGSN. GTP-U is used for carrying user data within EPC and the GPRS core network, and between the radio access network and the core network. The user data that is transported can be packets in any of IPv4, IPv6, or PPP formats. Different GTP variants are implemented by eNBs, S-GW, P-GWs, RNCs, SGSNs, and GGSNs within 3GPP networks. The mobile stations are connected to the network without being aware of GTP. GTP can be used with UDP or transport control protocol (TCP) [11].

3.1.2 Protocol stack for S1 user-plane

The S1 user-plane interface (S1-U) is a standard reference point between the eNB and the S-GW. The S1-U interface provides non-guaranteed delivery of user-plane PDUs between the eNB and the S-GW. The user-plane protocol stack on the S1 interface is shown in Figure 3.7. The transport network layer is built on IP transport and GTP-U is used on top of UDP/IP to carry the user-plane PDUs between the eNB and the S-GW.

3.1.3 Protocol stack for S1 control-plane

The S1 control-plane interface (S1-MME) is a standard reference point between the eNB and the MME. The control-plane protocol stack of the S1 interface is shown in Figure 3.8. The transport network layer is built on IP transport, similar to the user-plane, but for the reliable transport of signaling messages, SCTP is

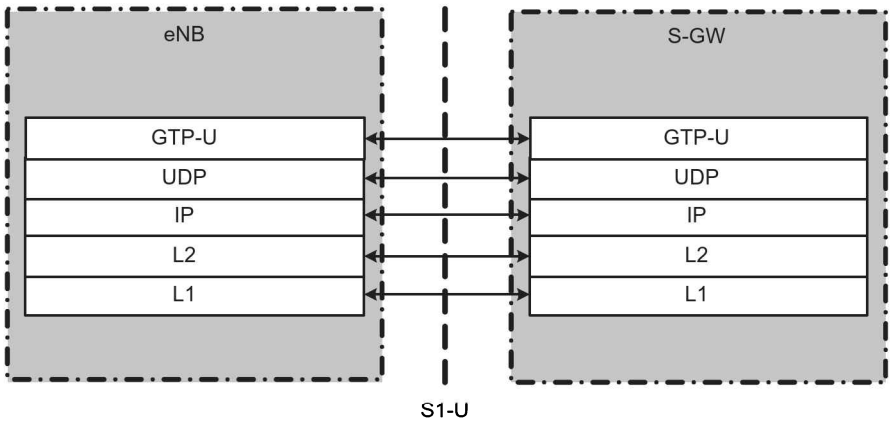


FIGURE 3.7

S1-U protocol architecture [1].

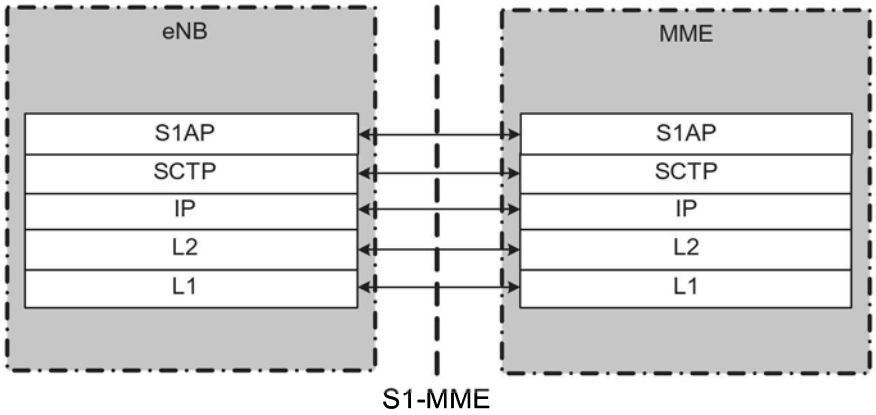


FIGURE 3.8

S1-MME protocol structure [1].

added on top of the IP layer. The application layer signaling protocol is referred to as S1-AP.

The SCTP layer provides the guaranteed delivery of application layer messages. The IP layer point-to-point transmission is used to deliver the signaling PDUs. A single SCTP association per S1-MME interface instance is used with one pair of stream identifiers for S1-MME common procedures. Only a few pairs of stream identifiers should be used for S1-MME-dedicated procedures. MME communication context identifiers that are assigned by the MME for S1-MME-dedicated procedures and eNB communication context identifiers that are assigned by the eNB for S1-MME-dedicated procedures are used to distinguish

UE-specific S1-MME signaling transport bearers. The communication context identifiers are conveyed in the respective S1AP messages. If the S1 signaling transport layer notifies the S1AP layer that the signaling connection is broken, the MME changes the state of the UEs which used this signaling connection to the ECM-IDLE state and the eNB releases the RRC connection with those UEs. In the case of relay nodes (RNs), there is one S1 interface between the RN and the relay-enabled eNB, and one S1 interface between the relay-enabled eNB and each of the MMEs in the MME pool. The S1 interface carries non-UE-associated S1AP signaling between the RN and the relay-enabled eNB as well as UE-associated S1AP signaling for UEs connected to the RN. The S1 interface between the relay-enabled eNB and an MME carries non-UE-associated S1AP signaling and UE-associated S1AP signaling for UEs connected to the RN and for UEs connected to the relay-enabled eNB [1].

3.1.4 Protocol stack for S5/S8 interface

The S5/S8 interface is defined between the S-GW and the P-GW. The S5 interface is used in non-roaming scenarios where the S-GW is located in the home network, or in roaming scenarios with both S-GW and P-GW located in the visited network. The latter scenario is also referred to as local breakout. The S8 interface is the roaming variant of S5 and is used in roaming scenarios with the S-GW in the visited network and the P-GW in the home network. When the GTP variant of the protocol is used, the S5/S8 interface provides the functionality associated with creation/deletion/modification/change of bearers for an individual user connected to the EPS. These functions are performed on a per PDN connection for each terminal. The S-GW provides the local anchor for all bearers for a single terminal and manages them toward the P-GW. Figure 3.9 shows the protocol stack over S5/S8 interface. Chapter 2 provides more information on the PMIP variant of S5/S8 [15].

3.1.5 Protocol stack for S11 interface

The S11 reference point defines the interface between the MME and the S-GW. This interface exclusively uses GTP-C and for LTE access. Due to the separation of the control- and user-plane functions between MME and S-GW, the S11 interface is used to create a new session (i.e., to establish the necessary resources for the session), and then to manage those sessions (i.e., modify, delete, and change any sessions for a terminal for each PDN connection) that has established connection within EPS. The S11 interface is always triggered by some events either directly from the NAS-level signaling from the terminal, such as a device attaching to the EPS network, adding new bearers to an existing session or creating a connection toward a new PDN, handover scenarios, or it may be triggered during network-initiated procedures such as P-GW-initiated bearer modification procedures. As such, the S11 interface coordinates the control- and user-plane procedures for a terminal during the period that the terminal is considered

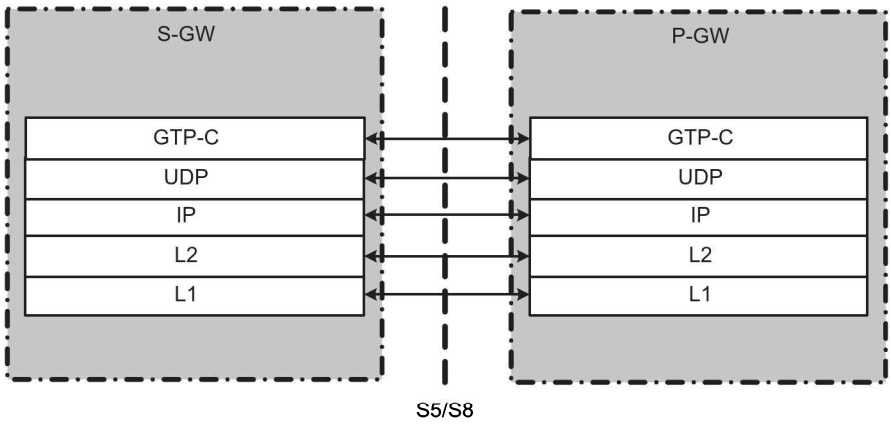


FIGURE 3.9

S5/S8 protocol structure [10].

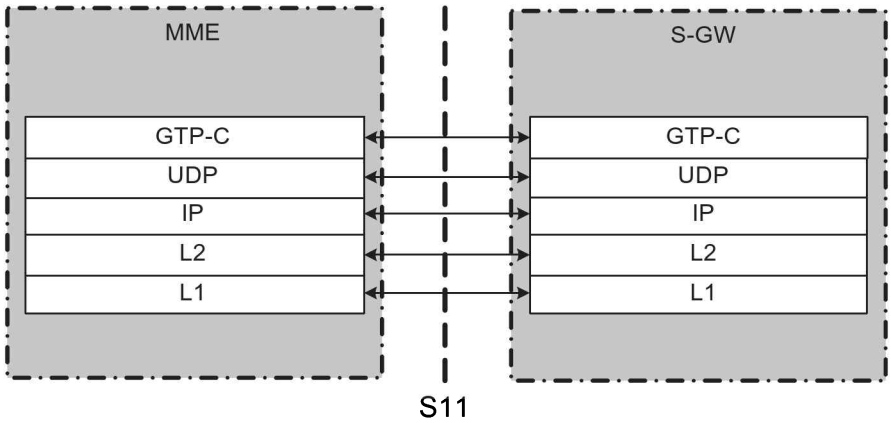


FIGURE 3.10

S11 protocol structure [15].

active/attached in the EPS. In the case of handover, the S11 interface is used to relocate the S-GW when appropriate, establish direct or indirect forwarding tunnel for user-plane traffic, and manage the user data traffic flow. Figure 3.10 shows the protocol stack across the S11 interface.

3.1.6 Protocol stack for X2 user-plane

The X2 user-plane interface defines data/control transport between eNBs. The X2-U interface provides non-guaranteed delivery of user-plane PDUs (due to use

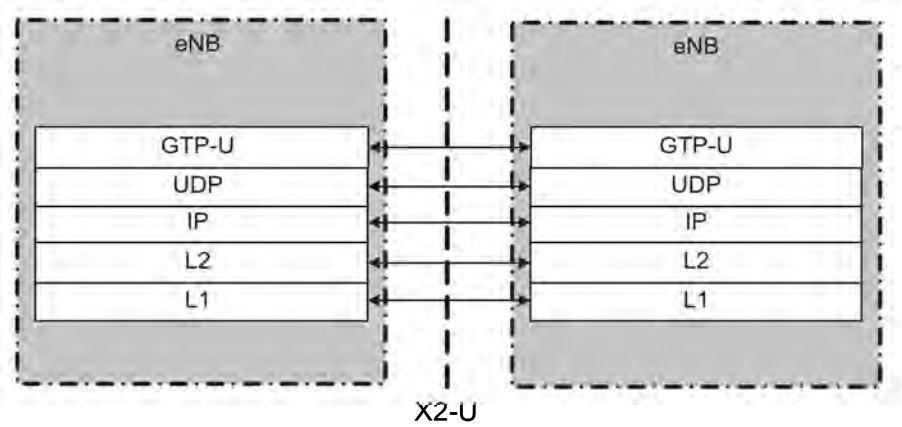


FIGURE 3.11

X2-U protocol structure [1].

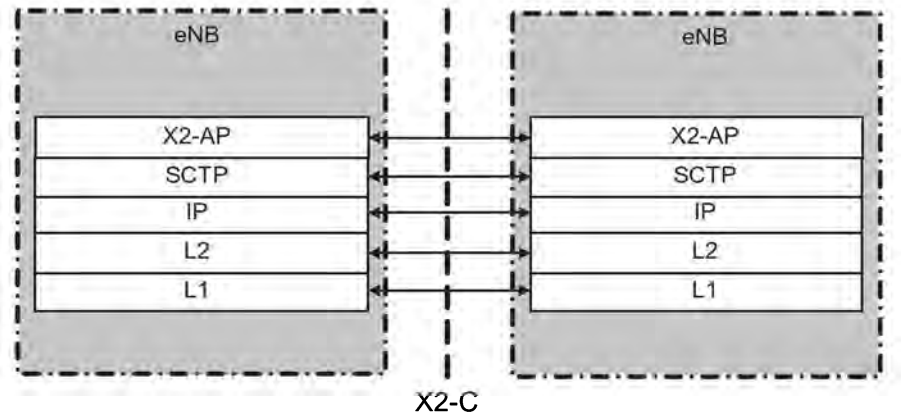


FIGURE 3.12

X2-C protocol structure [1].

of UDP transport protocol). The user-plane protocol stack on the X2 interface is shown in Figure 3.11. The transport layer is built on top of the IP layer and GTP-U is used on top of UDP/IP to carry the user-plane PDUs. The X2-U interface protocol stack is identical to the S1-U protocol stack [1].

The X2 reference point is a logical interface and even though it is normally shown as a direct connection between eNBs, it is usually routed via the same transport connection as the S1 interface to the site. The X2 interface control and user-plane protocol stacks are shown in Figures 3.11 and 3.12. X2 is an open interface and is often used for control-plane information exchange between neighboring eNBs; however, in the case of seamless handover, it can be temporarily used for

user data forwarding. The key difference between the user-plane and control-plane X2 interface protocol stacks is the use of SCTP for control-plane information transmission between eNBs. The use of SCTP enables reliable delivery of control-plane information between eNBs, while for data forwarding, the UDP is considered sufficient. The X2AP covers the radio-related signaling while the user-plane GTP is used for user data transport. The X2-AP functionalities are as follows [1,10]:

- Intra-LTE mobility management where the handover messages between eNBs are transmitted on the X2 interface.
- Load management to enable inter-cell interference coordination by providing information on the resource status, overload, and traffic condition between different eNBs.
- Setting up and resetting of the X2 interface.
- Error handling for covering specific or general error cases.

The X2 interface has a key role in the intra-LTE handover operation. The source-eNB uses the X2 interface to send the handover request message to the target-eNB. If the X2 interface does not exist between the serving and the target-eNBs, then procedures need to be initiated to set up one before handover can be initiated. The handover request message informs the target-eNB to reserve resources and to send the handover request acknowledgment message assuming resources are available.

3.1.7 Protocol stack for X2 control-plane

The X2 control-plane interface defines control and signaling transport between two neighbor eNBs. The control-plane protocol stack of the X2 interface is shown in Figure 3.12. The SCTP is used for reliable transport of control and signaling information. The application layer signaling protocol is referred to as X2-AP. A single SCTP association per X2-C interface instance is used with one pair of stream identifiers for X2-C common procedures. Only a few pairs of stream identifiers may be used for X2-C dedicated procedures. Source-eNB communication context identifiers that are assigned by the source-eNB for X2-C dedicated procedures, and target-eNB communication context identifiers that are assigned by the target-eNB for X2-C dedicated procedures, are used to distinguish UE-specific X2-C signaling transport bearers. The communication context identifiers are conveyed in the respective X2-AP messages.

Inter-cell interference coordination in E-UTRAN is performed through the X2 interface. When the interference conditions change, the eNB signals the new condition to its neighbor eNBs, e.g., the neighbor eNBs for which an X2 interface is configured due to mobility requirements. When a time-domain inter-cell interference coordination scheme (e.g., eICIC) is used to mitigate inter-cell interference, the reference eNB signals its almost blank subframe (ABS) patterns to the neighboring eNBs, so that the receiving eNBs can exploit the ABS pattern of the reference eNB and adapt their activities accordingly in order to minimize the inter-cell

interference. The load indication procedure is used to transfer interference coordination information between neighboring eNBs managing intra-frequency cells [1].

3.2 E-UTRAN and NAS protocols

The E-UTRAN overall architecture is described in 3GPP TS 36.300 [1,10]. Some general principles taken into consideration in the design of E-UTRAN architecture, as well as the E-UTRAN interfaces, are as follows:

- Signaling and data transport networks are logically separated.
- E-UTRAN and EPC functions are separated from transport functions. The addressing schemes used in E-UTRAN and EPC are not associated with the addressing schemes of transport functions. Some E-UTRAN or EPC functions reside in the same equipment.
- The mobility for RRC connection is controlled by the E-UTRAN.
- The interfaces are based on the logical model of the entity which is controlled through this interface.
- One physical network element can implement multiple logical nodes.

The E-UTRA radio access network consists of a network of eNBs, as illustrated in Figure 3.13. For unicast type of user traffic, there is no centralized controller in E-UTRAN. The eNBs are interconnected through X2 interface and to the EPC by means of S1 interface. The protocols operating between the eNBs and

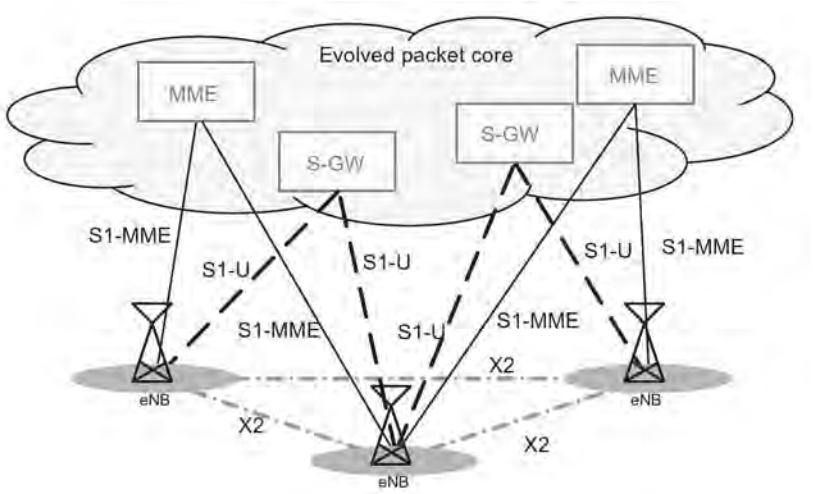


FIGURE 3.13

Radio access network entities and interfaces [1].

the UE are known as the AS protocols. The E-UTRAN is responsible for all radio-related functions, which can be summarized briefly as follows:

- Radio resource management which encompasses all functions related to the radio bearers, such as radio bearer control, admission control, mobility control, scheduling and dynamic allocation of resources to UEs in both uplink and downlink.
- Header compression which helps to ensure efficient use of the radio interface by compressing the IP packet headers that could otherwise cause a significant overhead, especially for small-packet applications such as VoIP.
- Security which ensures all data sent over the radio interface is encrypted.
- Connectivity to the EPC which consists of the signaling toward MME and the bearer path toward the S-GW.

On the network side, all of these functions reside in the eNBs, each of which can be responsible for managing multiple cells. Unlike the previous 3GPP releases, LTE integrates the radio controller function into the eNB. This allows tight interaction between different protocol layers of the radio access network, thus reducing latency and improving efficiency. Such distributed control eliminates the need for a complex central controller with high processing power, which in turn has the potential of reducing costs. Furthermore, since LTE does not support soft handovers similar to UMTS, there is no need for a centralized data-combining function in the network. One drawback for the lack of a centralized controller node is that, as the UE moves, the network must transfer the UE context, together with any buffered data, from one eNB to another. There are mechanisms and provisions in E-UTRA to avoid data loss during handover.

The general protocol model for E-UTRAN interfaces is depicted in Figure 3.14. The structure is based on the principle that the layers and planes are logically independent of each other. Therefore, as needed and when required, 3GPP can easily alter protocol stacks and planes to satisfy future requirements [1,10].

The protocol structure consists of two main layers, the radio network layer and transport network layer. E-UTRAN functions are realized in the radio network layer, and the transport network layer represents standard transport technology that is selected to be used for E-UTRAN. The control-plane includes the application protocol, i.e., S1-AP and X2-AP, and the signaling bearer for transporting the application protocol messages. The application protocol is used for setting up bearers (i.e., E-RAB[3]) in the radio network layer. The bearer parameters in the application protocol are not directly tied to the user-plane technology; rather, they are generic bearer parameters. The user-plane includes the data

[3]E-UTRAN radio access bearer (E-RAB) uniquely identifies the concatenation of an S1 bearer and the corresponding data radio bearer. When an E-RAB exists, there is a one-to-one mapping between the E-RAB and an EPS bearer of the non-access stratum as defined in 3GPP TS 24.301 [6].

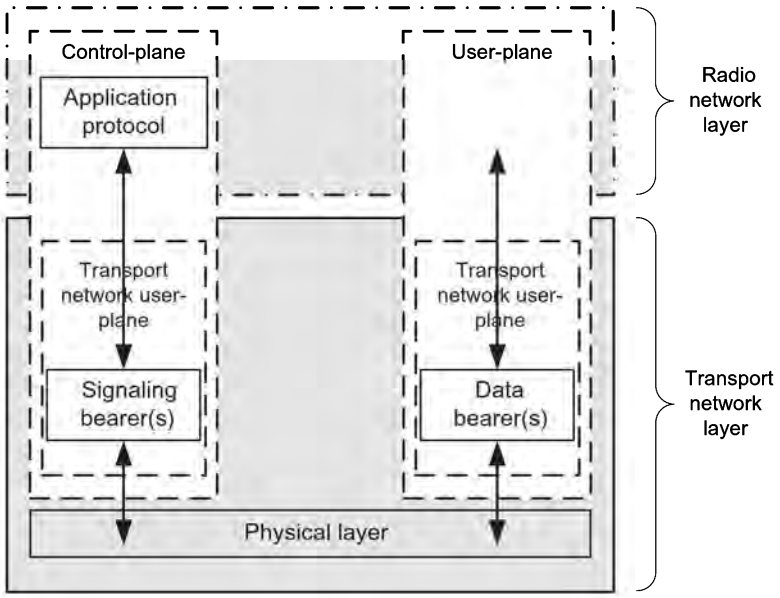


FIGURE 3.14

General protocol model for E-UTRAN interfaces [1].

bearer(s) for the data stream(s). The data stream is characterized by a tunneling protocol in the transport network layer.

As mentioned earlier, the eNB typically performs the following functions [1]:

- Radio resource management functions, including radio bearer control, radio admission control, connection management, dynamic allocation of resources to UEs in both uplink and downlink (i.e., scheduling).
- Header compression and encryption of user payloads.
- Selection of an MME at UE attachment when no routing to an MME can be determined from the information provided by the UE.
- Routing of user-plane data toward S-GW.
- Scheduling and transmission of paging messages (originated from the MME).
- Scheduling and transmission of broadcast information (originated from the MME).
- Measurement and reporting for support of mobility and scheduling.

As illustrated in Figure 3.13, the eNBs are interconnected with each other through an X2 interface. The X2 interface allows eNBs to communicate directly with each other and coordinate their activities. The X2 interface is split into separate control- and user-planes. The X2 control-plane carries X2AP messages between eNBs and uses SCTP for reliable delivery of messages. The X2AP is used to manage inter-eNB mobility and handovers, UE context transfers, inter-cell interference

management, and various error handling functions. The X2 user-plane uses GTP-U to tunnel user traffic between eNBs.

The main services and functions of the NAS sublayer include EPS bearer control, ECM-IDLE mobility handling, paging origination, and configuration and control of security [6]. The NAS procedures, and in particular the connection management procedures, are in essence similar to those of UMTS. The main change from UMTS is that EPS facilitates combination of some procedures to enable faster establishment of connections and bearers. The MME creates a UE context when a UE is turned on and attaches to the network. It assigns a unique short temporary identity to the UE that identifies the UE context in the MME. The UE context contains user subscription information obtained from the HSS. The local storage of subscription data in the MME allows faster execution of procedures such as bearer establishment. In addition, the UE context also holds dynamic information such as the list of bearers that are established and the terminal capabilities.

3.3 LTE-Uu interface protocols

The protocols over LTE-Uu and S1 interfaces are divided into two groups: user-plane protocols that are the protocols implementing the actual E-RAB service, i.e., carrying user data through the AS, and control-plane protocols, which are responsible for controlling the E-RABs and the connection between the UE and the network from different aspects including requesting the service, controlling different transmission resources, and handover. Figures 3.15 and 3.16 illustrate the LTE-Uu user-plane and control-plane protocol stacks, respectively. In the control-plane, the NAS functional block is used for network attachment, authentication, setting up bearers, and mobility management. All NAS messages are ciphered and integrity protected by the MME and the UE [1,5]. There is also a mechanism for transparent transfer of NAS messages that is included in the latter group. The E-RAB service is offered from service access point (SAP) to SAP by the AS.

The layer 2 functions in LTE are classified into MAC, RLC, and PDCP functions [1–5]. Figures 3.17 and 3.18 illustrate the structure of layer 2 protocols in LTE downlink and uplink. The SAP for between two adjacent protocol layers is marked with a circle at the interface between the sublayers. The SAP between the physical layer and the MAC sublayer provides the transport channels. The SAP between the MAC sublayer and the RLC sublayer provides the logical channels. The multiplexing of several logical channels on the same transport channel is performed by the MAC sublayer. Each logical channel is defined by the type of information that is transferred. The logical channels are generally classified into two groups: (1) control channels (for the transfer of control-plane information); and (2) traffic channels (for the transfer of user-plane information).

The RRC sublayer in the eNB, makes handover decisions based on neighbor cell measurements reported by the UE, performs paging of the users over the

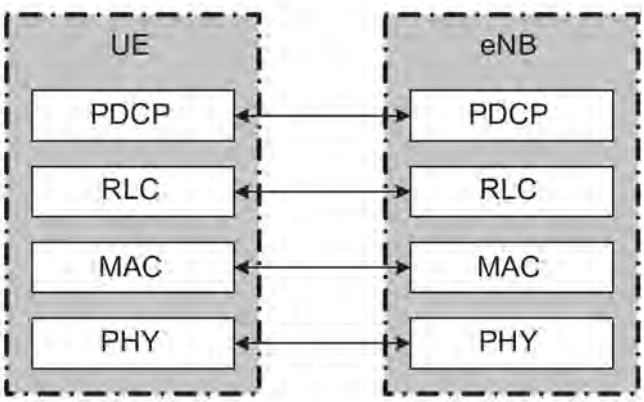


FIGURE 3.15

LTE-Uu interface user-plane protocol stack [1].

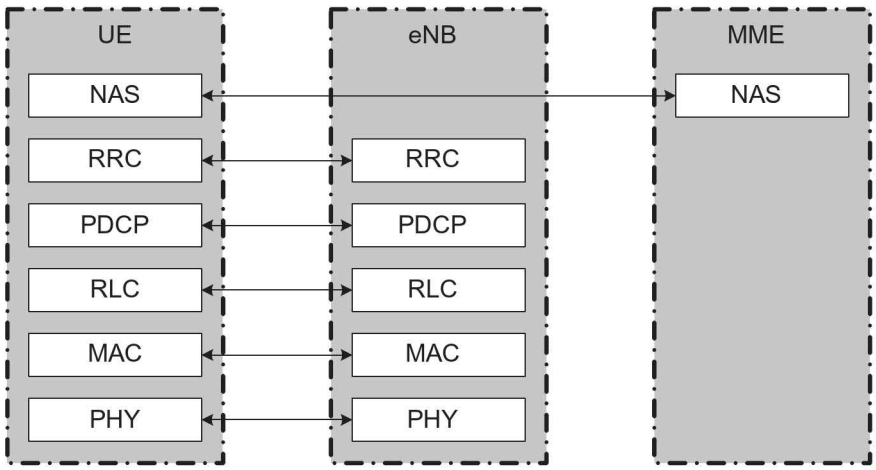


FIGURE 3.16

LTE-Uu interface control-plane protocol stack [1].

air-interface, broadcasts system information, controls UE measurement and reporting functions such as the periodicity of channel quality indicator reports, and further allocates cell-level temporary identifiers to active users. It also executes transfer of UE context from the serving eNB to the target-eNB during handover and performs integrity protection of RRC messages. The RRC sublayer is responsible for setting up and maintenance of radio bearers. Note that the RRC sublayer in 3GPP protocol hierarchy is considered as layer 3.

The services and functions provided by the PDCP sublayer in the user-plane include header compression and decompression of IP packets, transfer of user

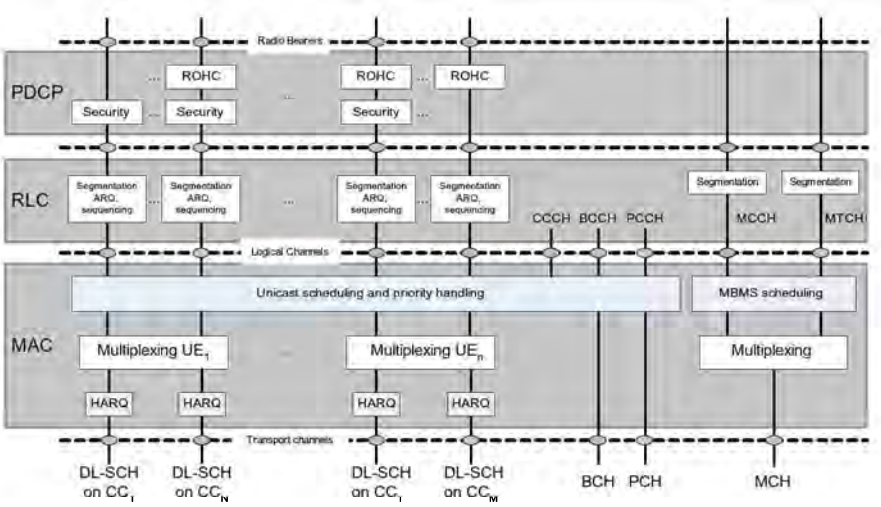


FIGURE 3.17

LTE layer 2 structure in the downlink [1].

data between NAS and RLC sublayer, sequential delivery of upper layer PDUs and duplicate detection of lower layer SDUs following a handover for RLC acknowledged mode, retransmission of PDCP SDUs following a handover for RLC acknowledged mode, and ciphering. The services and functions provided by the PDCP for the control-plane include ciphering and integrity protection and transfer of control-plane data where PDCP receives PDCP SDUs from RRC and forwards it to the RLC sublayer and vice versa [1,3,13].

The RLC sublayer is used to format and transport traffic between the UE and the eNB [1,5]. The RLC sublayer provides three different reliability modes for data transport, i.e., acknowledged mode (AM), unacknowledged mode (UM), and transparent mode (TM). The UM is suitable for transport of real-time services since such services are delay sensitive and cannot tolerate delay due to ARQ retransmissions. The acknowledged mode is appropriate for non-real-time services such as file transfers. The transparent mode is used when the size of SDUs are known in advance such as for broadcasting system information. The RLC sublayer also provides sequential delivery of SDUs to the upper layers and eliminates duplicate packets from being delivered to the upper layers. It may also segment the SDUs.

The MAC sublayer is responsible for data transfer and logical channel multiplexing, HARQ retransmissions and uplink/downlink scheduling, random access procedure, and maintenance of uplink timing. It is also responsible for multiplexing/de-multiplexing of data across multiple component carriers when carrier aggregation is used. From a physical layer perspective, the MAC sublayer provides and receives services in the form of transport channels. The data in a transport channel is organized into transport blocks. By varying the transport format of

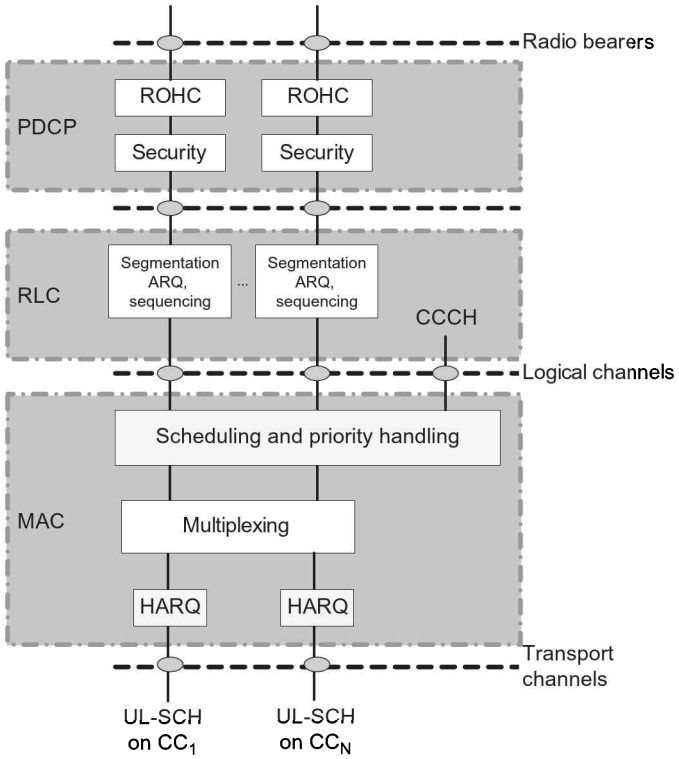


FIGURE 3.18

LTE layer 2 structure in the uplink [1].

the transport blocks, the MAC sublayer can realize different data rates and reliability levels. The MAC sublayer receives the RLC SDUs mapped to various logical channels in the downlink, and generates the MAC PDUs that further become the transport blocks in the physical layer.

The main difference between MAC and RRC control lies in the signaling reliability. The signaling corresponding to state transitions and radio bearer configurations should be performed by the RRC sublayer due to higher signaling reliability.

LTE-Advanced extended the capabilities of LTE Rel-8 with support of carrier aggregation, where two or more component carriers are aggregated in order to support wider transmission bandwidths up to 100 MHz. A user terminal may simultaneously receive or transmit one or multiple component carriers depending on its capabilities. From the UE perspective, the layer 2 aspects of HARQ are similar to those of LTE Rel-8. There is one transport block (in the absence of spatial multiplexing, up to two transport blocks in the case of spatial multiplexing) and one independent HARQ entity per scheduled component carrier. Each

transport block is mapped to a single component carrier on which all HARQ retransmissions may take place. A UE may be scheduled over multiple component carriers simultaneously, but at most one random access procedure is performed at any time. Whenever a UE is configured with only one component carrier, LTE Rel-9 DRX is the baseline. In other cases, the same DRX operation will be applied to all configured component carriers. Therefore, the layer 2 structure of the LTE-Advanced is similar to that of LTE Rel-8, except the addition of the carrier aggregation functionality; however, the multi-carrier nature of the physical layer is only exposed to the MAC sublayer through transport channels, where one HARQ entity is required per component carrier [1,4].

The logical and transport channel classification in E-UTRA are illustrated in Figures 3.19 and 3.20, respectively. Each logical channel is defined by the type of information that is transferred. The logical channels are generally classified into two groups: (1) control channels (for the transfer of control-plane information); and (2) traffic channels (for the transfer of user-plane information) as shown in Figure 3.19 [1]. The control channels are exclusively used for transfer of control-plane information. The control channels supported by MAC can be classified as follows:

- Broadcast control channel (BCCH): A downlink channel for broadcasting system information.
- Paging control channel (PCCH): A downlink channel that transfers paging information and system information change notifications. This channel is used for paging when the network does not know the location of the UE.

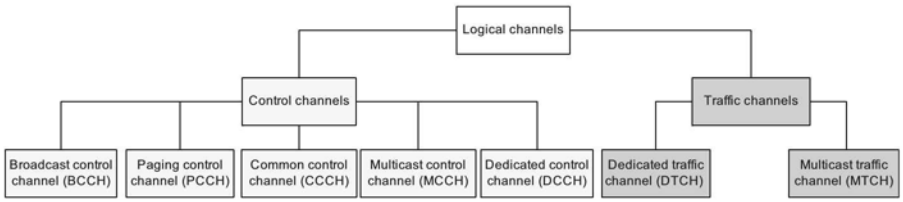


FIGURE 3.19

E-UTRA logical channel structure [1].

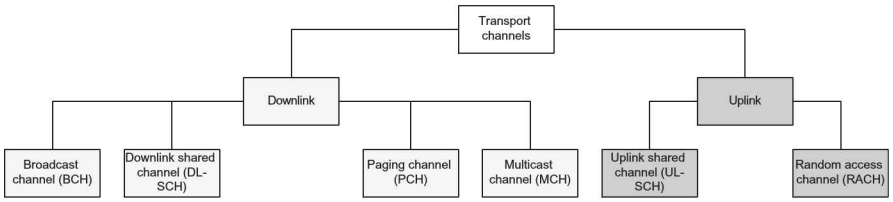


FIGURE 3.20

E-UTRA transport channel structure [1].

- Common control channel (CCCH): A channel for transmitting control information between UEs and eNBs. This channel is used for UEs that have no RRC connection with the network.
- Multicast control channel (MCCH): A point-to-multipoint downlink channel used for transmitting MBMS control information from the network to the UE for one or several MTCHs. This channel is only used by UEs that are subscribed to MBMS.
- Dedicated control channel (DCCH): A point-to-point bidirectional channel that transmits dedicated control information between a UE and the network. It is used by UEs that have RRC connection.

The traffic channels are exclusively used for the transfer of user-plane information. The traffic channels supported by MAC can be classified as follows (as shown in Figure 3.19):

- Dedicated traffic channel (DTCH): A point-to-point bi-directional channel dedicated to a single UE for the transfer of user information.
- Multicast traffic channel (MTCH): A point-to-multipoint downlink channel for transmitting broadcast/multicast data from the network to the UE. This channel is only used by UEs that receive MBMS.

The physical layer provides information transfer services to the MAC and higher layers. The physical layer transport services are described by how and with what characteristics data is transferred over the radio interface. This should be clearly distinguished from the classification of what is transported which relates to the concept of logical channels at the MAC sublayer. As shown in Figure 3.20, downlink transport channels can be classified as follows:

- Broadcast channel (BCH) is characterized by fixed, predefined transport format, and is required to be broadcast in the entire coverage area of the cell.
- Downlink shared channel (DL-SCH) is characterized by support for HARQ protocol, dynamic link adaptation by varying modulation, coding, and transmit power, possibility for broadcast in the entire cell, possibility to use beamforming, dynamic and semi-static resource allocation, UE DRX to enable power saving, and MBMS transmission.
- Paging channel (PCH) is characterized by support for UE DRX in order to enable power saving, requirement for broadcast in the entire coverage area of the cell and is mapped to physical resources which can also be used dynamically for traffic or other control channels.
- Multicast channel (MCH) is characterized by requirement to be broadcast in the entire coverage area of the cell, support for macro-diversity combining of MBMS transmission on multiple cells, and support for semi-static resource allocation.

The uplink transport channels are classified as follows (see Figure 3.20):

- Uplink shared channel (UL-SCH) is characterized by possibility to use beamforming, support for dynamic link adaptation by varying the transmit power

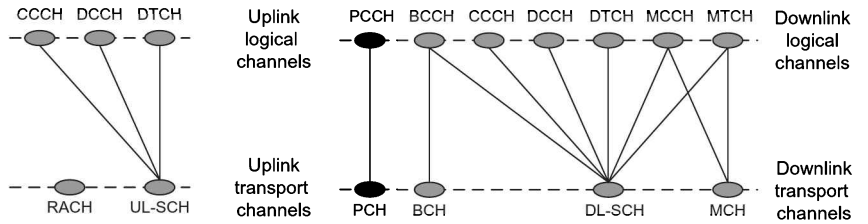


FIGURE 3.21

Logical to transport channel mapping [1].

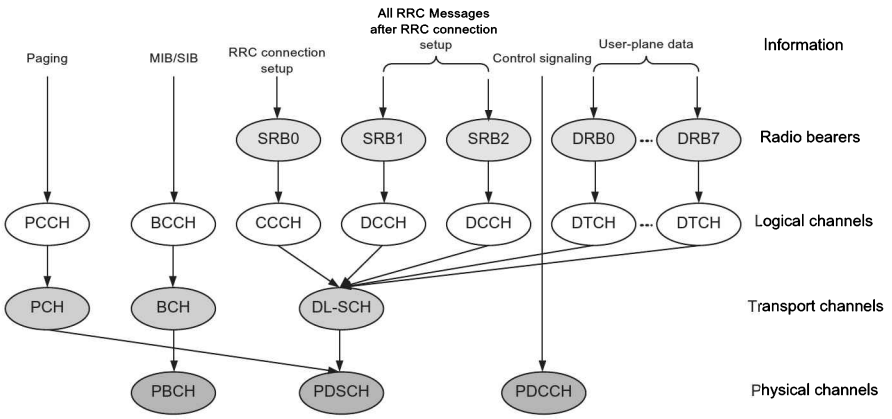


FIGURE 3.22

Mapping of downlink information to radio and signaling bearers as well as various logical, transport, and physical channels [1–5].

and modulation and coding schemes, support for HARQ, support for both dynamic and semi-static resource allocation.

- Random access channel (RACH) is characterized by limited control information and collision risk.

The mapping of the logical channels to the transport channels in the downlink and uplink is shown in Figure 3.21.

The mapping of downlink and uplink information to radio and signaling bearers as well as various logical, transport, and physical channels at different layers of protocol stack are illustrated in Figures 3.22 and 3.23.

3.4 Support for relay nodes

The E-UTRAN architecture was modified in LTE Rel-10 to include support for relaying function. The intermediate relay nodes are connected to a serving eNB

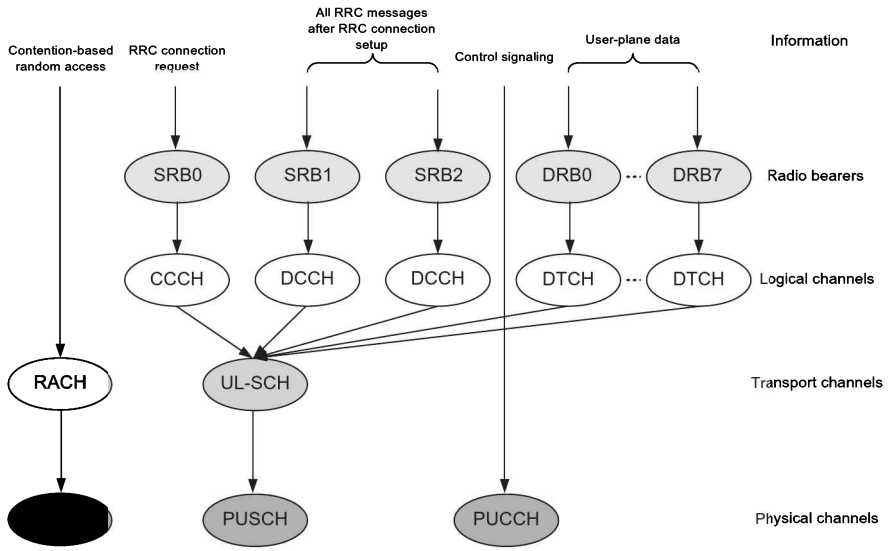


FIGURE 3.23

Mapping of uplink information to radio and signaling bearers as well as various logical, transport, and physical channels [15].

called the donor eNB (DeNB), via a modified version of the E-UTRA radio interface known as Un interface. The relay entity supports the eNB functionality, i.e., it terminates the radio protocols of the E-UTRA radio interface and the S1 and X2 interfaces. In addition to the eNB functionality, the relay node (RN) also supports a subset of the UE functionality, e.g., physical layer, layer 2, RRC, and NAS functionality, in order to wirelessly connect to the DeNB.

The modified architecture for supporting the relay nodes is shown in Figure 3.24. The RN terminates S1, X2, and Un interfaces. The DeNB provides S1 and X2 proxy functionality between the RN and other eNBs, MMEs, and S-GWs. The S1 and X2 proxy functionality includes forwarding UE-dedicated S1 and X2 signaling messages as well as GTP data packets between the S1 and X2 interfaces associated with the RN and the S1 and X2 interfaces associated with other network nodes. Due to the proxy functionality, the DeNB appears as MME (for S1-MME), eNB (for X2), and S-GW (for S1-U) to the RN. The DeNB can provide the S-GW/P-GW type functions needed for the RN operation. This includes creating a session for the RN and managing EPS bearers for the RN, as well as terminating the S11 interface toward the MME serving the RN. The RN and DeNB map signaling and data packets to EPS bearers that are set up for the RN. The mapping is based on the existing QoS mechanisms defined for the UE and the P-GW. The P-GW functions in the DeNB can allocate an IP address for the RN for the purpose of operation and management, which may be different from the S1 IP address of the DeNB.

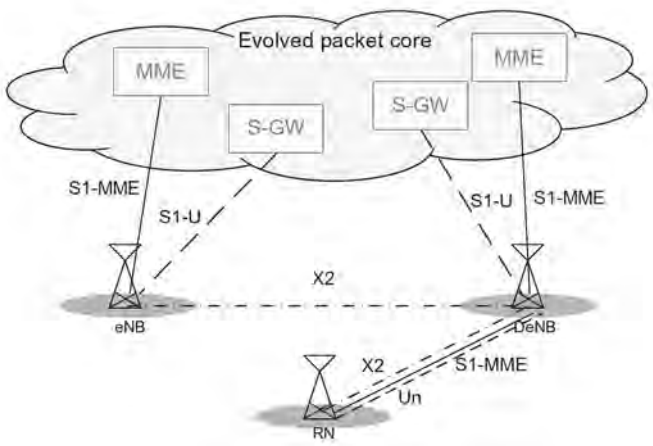


FIGURE 3.24

Modified E-UTRAN architecture supporting relay nodes [1,9].

The S1 user-plane protocol stack for supporting RNs is shown in Figure 3.25. There is a GTP tunnel associated with each UE EPS bearer, connecting the S-GW associated with the UE to the DeNB, which is switched to another GTP tunnel in the DeNB, going from the DeNB to the RN (one-to-one mapping). The X2 user-plane protocol stack for supporting RNs is also shown in Figure 3.25. There is a GTP forwarding tunnel associated with each UE EPS bearer subject to forwarding, connecting the other eNBs to the DeNB, which is switched to another GTP tunnel in the DeNB, going from the DeNB to the RN (one-to-one mapping). The S1 and X2 user-plane packets are mapped to radio bearers over the Un interface. The mapping can be based on the QoS class identifier associated with the UE EPS bearers. The UE EPS bearer with similar QoS can be mapped to the same Un radio bearer [1,9].

The S1 control-plane protocol stack for supporting RNs is shown in Figure 3.26. There is a single S1 interface between each RN and the DeNB, and there is a single S1 interface between the DeNB and the MME. The DeNB processes and forwards all S1 messages between the RN and the MMEs for all UE-dedicated procedures. The processing of S1-AP messages includes modifying S1-AP UE identifiers, transport-layer address, and GTP tunnel endpoint identifiers (TEIDs) without changing other parts of the messages. All non-UE-dedicated S1-AP procedures are terminated at the DeNB and are handled locally between the RN and the DeNB, and between the DeNB and the MME(s). Upon reception of an S1 non-UE-dedicated message from an MME, the DeNB may trigger corresponding S1 non-UE-dedicated procedure(s) to the RN(s). If more than one RN is involved, the DeNB may wait and aggregate the response messages from all participating RNs before responding to the MME. Upon reception of an S1 non-UE-dedicated message from an RN, the DeNB may trigger associated S1 non-UE-dedicated procedure(s) to the MME(s).

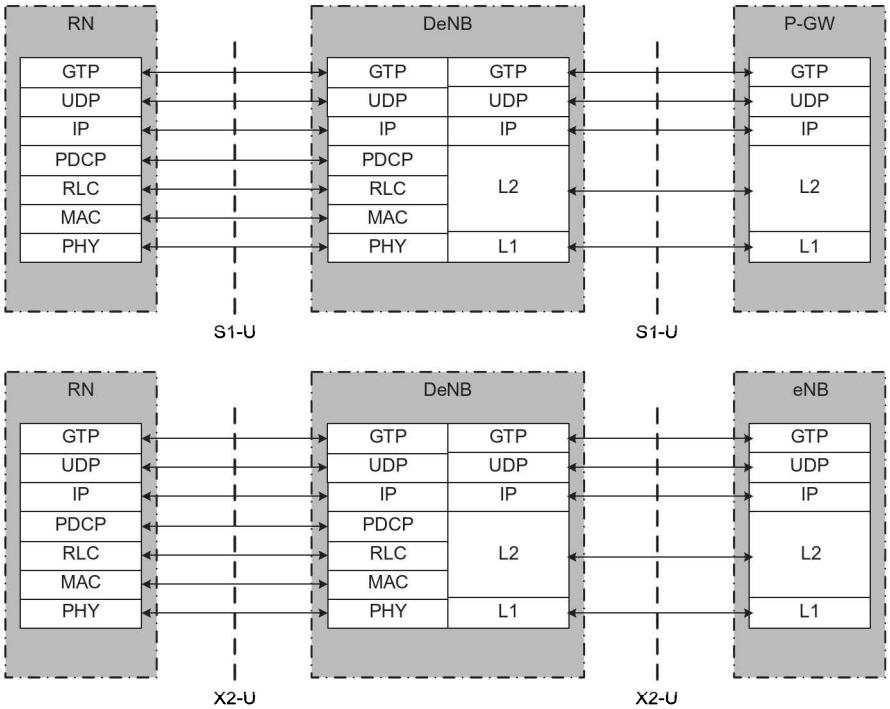


FIGURE 3.25

Modified S1-U and X2-U protocol structure for support of relay nodes [1].

The X2 control-plane protocol stack for supporting RNs is shown in Figure 3.26. There is a single X2 interface between each RN and the corresponding DeNB. The DeNB may have additional X2 interfaces to the neighboring eNBs. The DeNB processes and forwards all X2 messages between the RN and other eNBs for all UE-dedicated procedures. The processing of X2AP messages includes modifying S1AP or X2AP UE identifiers transport-layer address, and GTP TEIDs without changing other parts of the message. All non-UE-dedicated X2AP procedures are terminated at the DeNB and are handled locally between the RN and the corresponding DeNB, or between the DeNB and other eNBs. If one or more RNs are involved, the DeNB may wait and aggregate the response messages from all participating RNs to respond to the originating node [1,9].

The RN connects to the DeNB via a Un interface using the same radio protocols and procedures used by a UE to connect to an eNB. In addition to UE-type protocol functionalities, the RRC sublayer is able to configure and reconfigure an RN subframe configuration through the RN reconfiguration procedure over the Un interface, e.g., downlink subframe configuration and an RN-specific control channel for transmissions between the RN and the DeNB. The RN may request such a configuration from the DeNB during the RRC

Samsung Ex. 1005, Page 178 of 1147

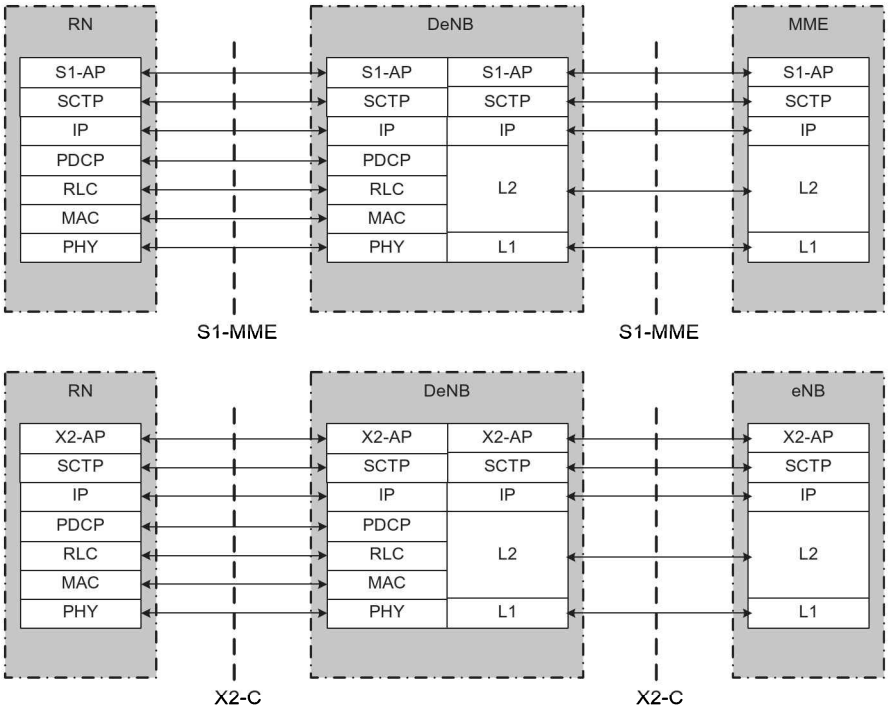


FIGURE 3.26

Modified S1-C and X2-C protocol structure for support of relay nodes [1].

connection establishment, and the DeNB may initiate the RRC signaling for such a configuration. The RN subframe configuration over the Un interface may be temporarily misaligned with the MBSFN subframes configured in the relay cell because a new subframe configuration might have been applied earlier by the RN on the Un reference point. The RRC sublayer of the Un interface can further send updated system information in a dedicated message to an RN with an RN subframe configuration. The PDCP sublayer of the Un interface can provide integrity protection for the user-plane. The integrity protection is configured per data radio bearer [1,9].

3.5 Support for HeNB nodes

Small cells are low-power wireless access points that operate in licensed spectrum and are operator-managed. Small cells can provide improved cellular coverage, and capacity, which have applications for homes and enterprises as well as metropolitan and rural public spaces. Various types of small cells include femto-cells, pico-cells, metro-cells, and micro-cells—broadly increasing in size from femto-cells (the

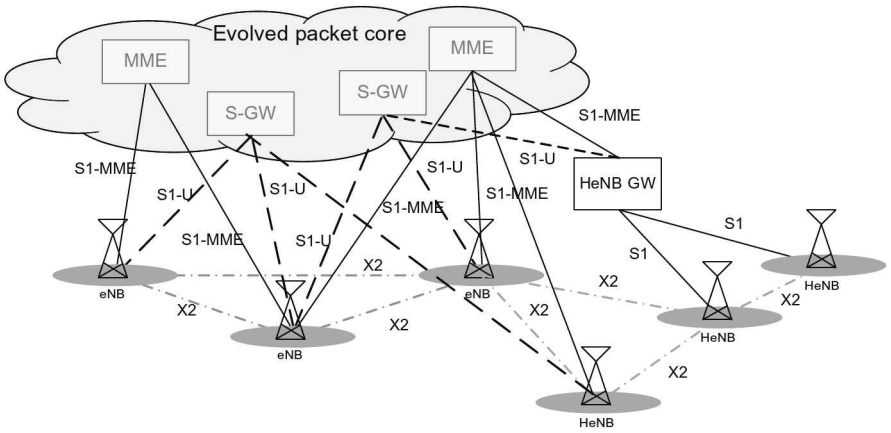


FIGURE 3.27

Modified E-UTRAN architecture with support of HeNB-GW and HeNB [1].

smallest) to micro-cells (the largest). Any or all of these small cells can be based on femto-cell technology, referred to as home eNB (HeNB) in 3GPP literature. Small cells also facilitate a new breed of mobile service that exploits the technology's ability to detect presence and connect and interact with the existing networks.

As shown in Figure 3.27, the E-UTRAN architecture may be extended to include a home eNB gateway (HeNB-GW) to provide an S1 interface between the HeNB entities (also known as femto-cell in the literature) and the EPC [1]. The HeNB-GW serves as a concentrator for the control-plane and the S1-MME interface. The S1-U interface from the HeNB may be terminated at the HeNB-GW, or a direct logical user-plane connection between the HeNB and S-GW may be used. In the modified architecture, the S1 interface is defined between the HeNB-GW and the core network, between the HeNB and the HeNB-GW, between the HeNB and the core network, or between the eNB and the core network. The HeNB-GW would appear to the MME as an eNB and to the HeNB as an MME. The S1 interface between the HeNB and the EPC is the same, regardless of the way HeNB is connected to the EPC.

The HeNB-GW is connected to the EPC in such a way that inbound and outbound mobility to cells served by the HeNB-GW does not require inter-MME handovers. Each HeNB serves only one cell. The functions and protocols supported by the HeNB are the same as those supported by an eNB, with the exception that a new reference point S5 may be used between the HeNB and the S-GW, if local IP access is supported. LTE Rel-11 supports direct X2-connectivity between HeNBs, irrespective of the HeNB-GW existence [1].

If the HeNB supports the local IP access feature, then it is required to support the S5 interface with the S-GW and the SGi interface with the IP network. As mentioned earlier, a HeNB performs the same functions as an eNB with some additional functions and restrictions in the case of connection to the HeNB-GW,

Samsung Ex. 1005, Page 180 of 1147

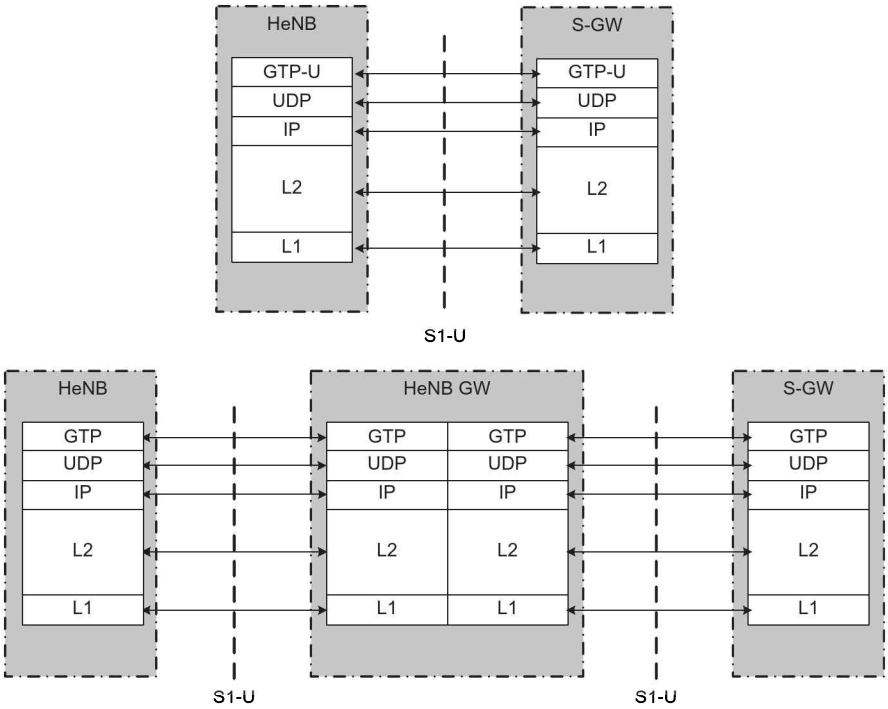


FIGURE 3.28

S1-U interface between HeNB and S-GW with and without HeNB-GW [1].

such as discovery of a suitable serving HeNB-GW, connection to a single HeNB-GW at any time, using the same tracking area code and PLMN ID supported by the HeNB-GW, etc. HeNBs may be deployed without network planning. A HeNB may be moved from one geographical area to another, and therefore it may need to connect to different HeNB-GWs depending on its location. The HeNB may support fixed broadband access network interworking function to signal tunnel information to the MME. The tunnel information includes the HeNB IP address.

The HeNB-GW performs the following functions [1]:

- Relaying UE-associated S1 application messages between the MME serving the UE and the HeNB serving the UE, except the UE context release request message received from the HeNB.
- Terminating non-UE-associated S1 application procedures in the direction of the HeNB or the MME.
- Terminating the S1-U interface with the HeNB and with the S-GW.
- Supporting tracking area code and PLMN ID used by the HeNB.
- Routing the S1 path-switch request message toward the MME based on the globally unique MME identifier of the source MME received from the HeNB.

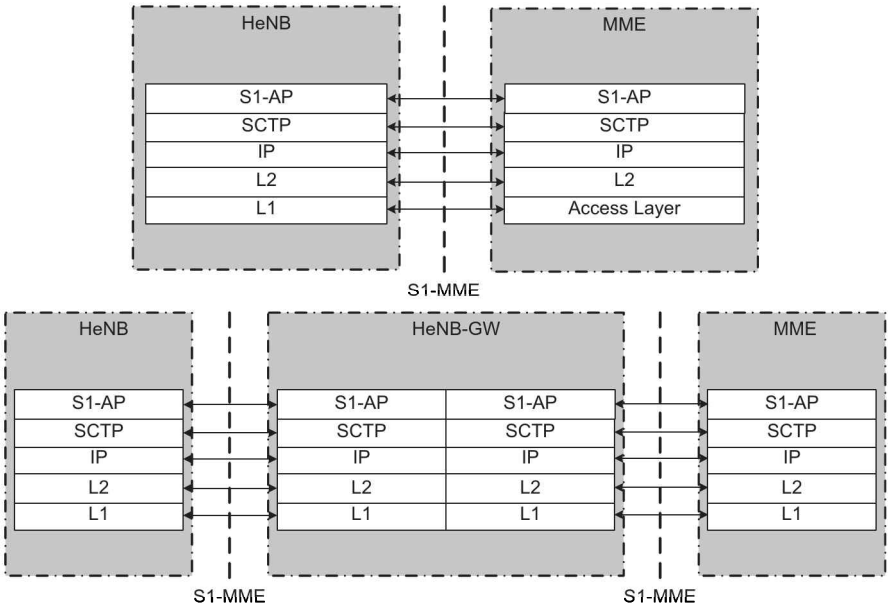


FIGURE 3.29

S1-MME interface between HeNB and MME with and without HeNB-GW [1].

As shown in Figure 3.28, the user-plane interface between the HeNB, HeNB-GW, and the S-GW can be defined with and without the HeNB-GW.

The HeNB-GW may optionally terminate the user-plane toward the HeNB or the S-GW, and forward the user-plane data between the HeNB and the S-GW. As shown in Figure 3.29, the control-plane interface between the HeNB and MME can be defined irrespective of the HeNB-GW presence. When the HeNB GW is not present, all S1AP procedures are terminated at the HeNB and the MME. The HeNB-GW forwards control-plane information between the HeNB and the MME.

References

3GPP Specifications

[1] 3GPP TS 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 11), December 2012.

[2] 3GPP TS 36.331, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 11), December 2012.

[3] 3GPP TS 36.323, Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) specification (Release 11), December 2012.

[4] 3GPP TS 36.321, Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification (Release 11), December 2012.
[5] 3GPP TS 36.322, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification (Release 11), December 2012.
[6] 3GPP TS 24.301, Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3 (Release 11), December 2012.
[7] 3GPP TS 29.060, General Packet Radio Service (GPRS); GPRS Tunneling Protocol (GTP) across the Gn and Gp interface (Release 11), December 2012.
[8] 3GPP TS 23.203, Policy and charging control architecture (Release 11), December 2012.
[9] 3GPP TR 36.806, Evolved Universal Terrestrial Radio Access (E-UTRA); Relay architectures for E-UTRA (LTE-Advanced) (Release 9), April 2010.
[10] 3GPP TS 36.401, Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Architecture description (Release 11), December 2012.

IETF Specifications

[11] IETF RFC 768, J. Postel, User Datagram Protocol, August 1980.
[12] IETF RFC 2460, S. Deering, Internet Protocol, Version 6 (IPv6) Specification, December 1998.
[13] IETF RFC 3095, C. Bormann, RObust Header Compression (ROHC): Framework and four profiles: RTP, UDP, ESP, and uncompressed, July 2001.
[14] IETF RFC 4960, R. Stewart, Stream Control Transmission Protocol, September 2007.

Books

[15] Magnus Olsson, et al., SAE and the Evolved Packet Core: Driving the Mobile Broadband Revolution, Academic Press, 2009.
[16] A.S. Tanenbaum, Computer Networks, fifth ed., Prentice Hall, 2010.

Articles

[17] Open System Interconnection Reference Model, Wikipedia, <http://en.wikipedia.org/wiki/OSI_model>.
[18] White paper, The LTE network architecture: A comprehensive tutorial, Alcatel Lucent, 2009.
[19] ITU-T Recommendation X.200: Open Systems Interconnection, Basic Reference Model: The basic model, July 1994.

CHAPTER

System Operation and UE States

4

CHAPTER CONTENTS

4.1 General UE behavior 154
4.2 UE initialization procedure (cell search and cell selection) 157
4.3 EPS connection and mobility management 166
4.3.1 EPS mobility management 167
4.3.2 EPS connection management 168
4.4 UE states 169
4.4.1 Idle mode procedures 171
4.4.1.1 Tracking area update 174
4.4.1.2 Paging 174
4.4.1.3 Cell reselection 176
4.4.1.4 Idle mode DRX 177
4.4.2 Connected mode procedures 178
4.5 Network attach procedures 180
4.5.1 Random access procedure 185
4.5.2 RRC connection establishment, modification, and release 188
4.5.3 Service establishment and QoS 191
4.6 Data call establishment procedures 194
4.6.1 UE-originated data call procedure 194
4.6.2 UE-terminated data call procedure 196
4.7 Mobility procedures 198
4.7.1 Idle mode mobility procedures 201
4.7.2 Connected mode mobility procedures 202
4.7.3 HetNet mobility 203
4.7.4 S1-based handover procedure 208
4.7.5 X2-based handover without S-GW change 212
4.7.6 X2-based handover with S-GW change 214
4.7.7 Inter-RAT mobility procedures 216
4.8 Overview of IP packet processing procedures 222
References 224

LTE-Advanced.

© 2014 Elsevier Inc. All rights reserved.

Samsung Ex. 1005, Page 184 of 1147

This chapter provides an in-depth systems view of the fundamental procedures, signaling, and protocols utilized by a mobile terminal for network discovery; cell selection/cell reselection; network attach/detach; authentication and security establishment; capability negotiation; signaling related to control-plane and user-plane connection establishment (access stratum (AS) and non-access stratum (NAS)); and IP connection establishment using state diagrams and signal flow graphs. We will also discuss radio resource management, paging, UE tracking, mobility, and power management functions. We will further explain practical considerations and real-life issues of network and terminal operation followed by references for additional reading. In this chapter, we will only focus on the procedures and will leave the description of the details of functional processing at different stages to the next chapters. An attempt has been made to characterize the behavior of the mobile terminal and the E-UTRAN in various operating conditions and states including idle mode and connected mode procedures, EPS connection and mobility management states, service establishment and QoS control, as well as new topics including mobility in heterogeneous networks.

State diagrams are used to describe the behavior of a system. They can describe possible states of a system and transitions between them as certain events occur or based on the system inputs. The system described by a state diagram must be composed of a finite number of states. However, in some cases, the state diagram may represent a reasonable abstraction of the system. There are many forms of state diagrams which differ slightly and have different semantics. State diagrams can be used to graphically represent finite state machines (i.e., a model of behavior composed of a finite number of states, transitions between those states, and actions). A state is defined as a finite set of procedures or functions that are executed in a unique order. In the state diagram, each state may have some inputs and outputs, where deterministic transitions to other states or the same state happen based on certain conditions. In this chapter, the notion of mode is used to describe a substate or a collection of procedures/protocols that are associated with a certain state. The unique definition of states and their corresponding modes and protocols, and internal and external transitions, is imperative to the unambiguous behavior of the system. Also, it is important to show the reaction of the system to unsuccessful execution of certain procedures.

4.1 General UE behavior

When a UE is powered on or upon loss of coverage, it needs to select an appropriate PLMN and radio access technology. The main input to PLMN and RAT selection procedures is provided by the user in the form of preferred RAT, or is obtained from the universal subscriber identity module (USIM) card in the form

of the most recently used PLMN. In general, a UE at the power-on stage has some information and settings, and requires some other information in order to select the network and attach to it. The information that the UE obtains from its USIM includes the network operator's PLMN list and subscription information. In addition to the information on the USIM, the UE has some stored information in its memory (additional knowledge), which includes most recently used frequency band and PLMN; tracking area code (TAC); cell identifier; SAE temporary mobile subscriber identity (S-TMSI) [1]; and inter-RAT frequency band. The UE capability determines what the UE is capable of in terms of frequency bands, duplexing schemes, multi-antenna schemes, data buffering capacities, etc. This information includes the UE category, frequency band, synchronization signal sequences, general radio resource information, multiantenna mode parameters (e.g., codebooks), duplex mode, and preamble sequence generation algorithm.

In order to perform cell selection, the UE needs to obtain the frequency and timing synchronization information, system bandwidth, number of eNB transmit antennas, cell IDs (e.g., cell radio network temporary identifier (C-RNTI), physical cell ID, TAC), PLMN ID, signaling/data radio resource information, RACH_ROOT_SEQUENCE, and physical random access channel configuration to perform a random access procedure.

During cell selection, the UE searches for a suitable cell in the selected PLMN and RAT, and chooses that cell in order to initially gain access to the available services by tuning to its control channels, i.e., the UE camps on that cell. The UE then registers in the tracking area of the selected cell and RAT by performing the attach procedure. The outcome of this process is the successful location registration where the selected PLMN becomes the registered PLMN. The UE now continuously searches for the best cell on which to camp. The particular cell reselection process depends on the cell reselection scenario, UE capabilities (e.g., multi-RAT support, multi-band support), subscriber priority class, etc.

While the baseline for general UE behavior in LTE is inherited from GSM and UMTS, an LTE-specific aspect of the UE behavior is handling of priorities between available RATs and frequency layers, where a set of broadcast system information parameters configure a default RAT and frequency layer priority list. The RRC signaling is used to configure a UE-specific priority list [9]. The priority list provides the network with the means to direct the UEs to camp on particular RATs and frequency layers, facilitating load balancing functions. In the above priority scheme, inter-frequency and inter-RAT cell reselection relies on two sets of (cell-specific and UE-specific) parameters. The cell-specific set of parameters are common to all UEs camping on a specific RAT and frequency, whereas the UE-specific set of parameters allow the operators to control the UE behavior depending on UE class and operator-defined policies.

Figure 4.1 illustrates the state transitions of a UE during normal operation. The figure shows the transition from power-on to RRC connection setup and establishment of user-plane and control-plane with the serving cell; RRC connection release and transition to RRC_IDLE state; cell reselection and transition to another cell; handover to another cell while in RRC_CONNECTED state; and cell redirection from one cell to another upon RRC connection release. We will discuss cell selection, cell reselection, handover, and cell redirection procedures as well as other aspects of LTE system operation in the idle and connected modes in the next sections.

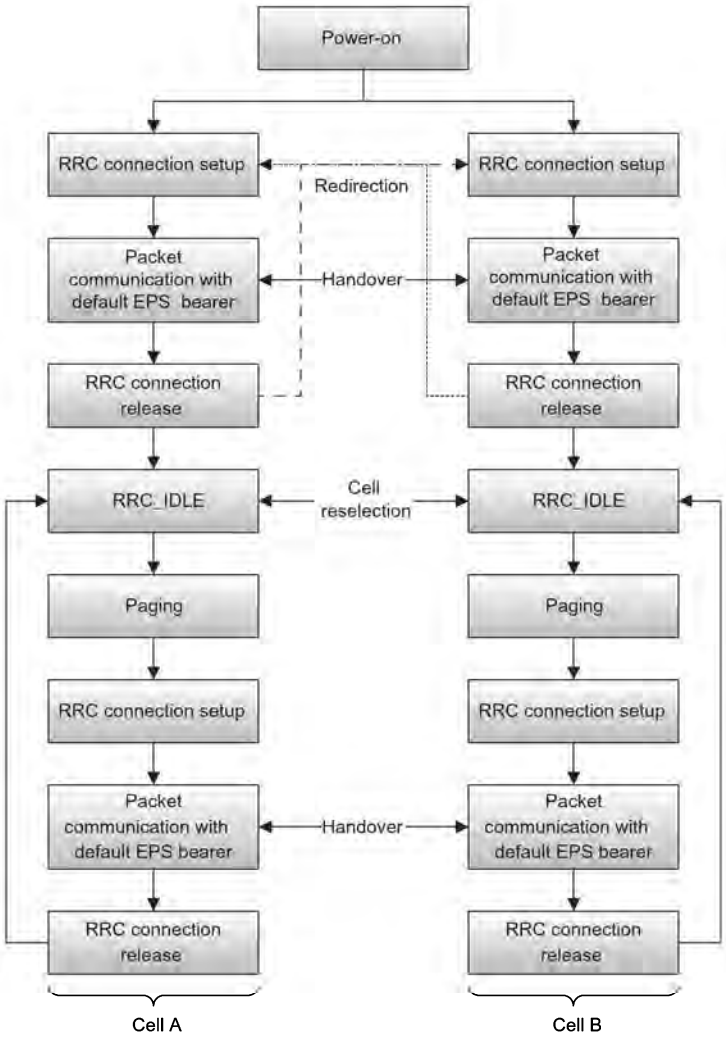


FIGURE 4.1

General UE state machine [32].

4.2 UE initialization procedure (cell search and cell selection)

The UE's USIM card stores the network operator's PLMN ID list and subscription information. The subscription information defines the policy for system selection and helps a UE to acquire a particular network. This feature enables, for example, an existing 3G network operator to deploy LTE by configuring the policy in order to initially instruct the UE to select the LTE network. In addition to the above information, the UE also knows the list of the frequency bands supported by LTE and other RATs such as GSM/UMTS, the most recently used LTE frequency band(s), PLMN IDs, TACs, cell IDs, S-TMSI, and inter-RAT frequency bands, as well as the UE category, synchronization signal sequences, general radio resource information, general MIMO parameters such as codebooks, duplex modes, and preamble sequence generation algorithm. However, the UE needs to acquire network frequency and timing synchronization information, system bandwidth, number of eNB transmit antennas, C-RNTI, cell ID, TAC, RACH_ROOT_SEQUENCE, specific radio resources used for control signaling, and PRACH configuration in order to be able to connect to a particular LTE network. In certain cases, the UE might have experienced link failure and has to make an attempt to reacquire the network using the most recent information that it had used earlier.

As shown in Figure 4.2, in order to select a PLMN, the UE must perform a cell search procedure in the supported frequency bands. The procedure requires

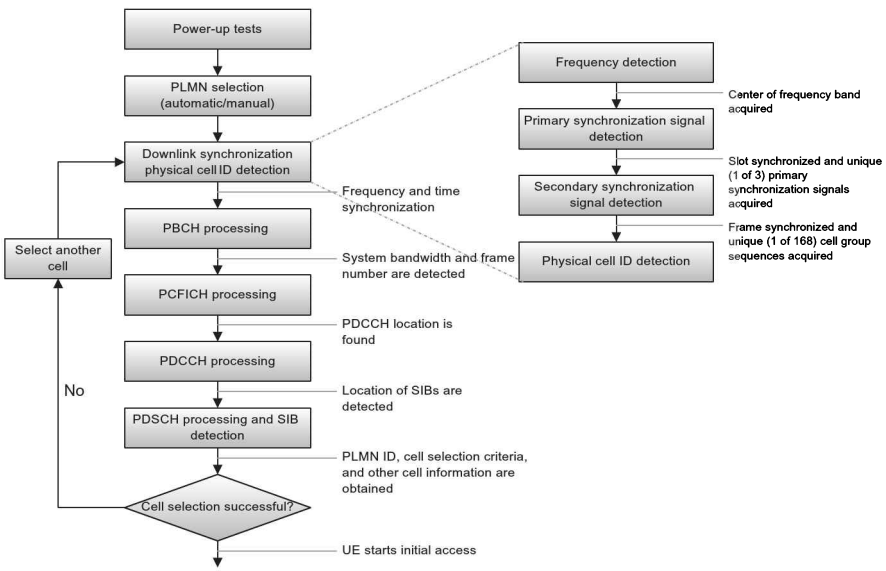


FIGURE 4.2

UE initial cell search and selection [20–22,33].

that the UE attain time and frequency synchronization with the specific cell. This enables decoding of the PBCH which carries the master information block (MIB) that contains the critical system information necessary to decode transmissions on a PDSCH. System information blocks (SIBs) are scheduled for transmission on PDSCH via SI messages. System information block type 1 (SIB1) contains a list of PLMN IDs along with a flag to show if a specific PLMN is reserved for operator use.

The selection of a PLMN can be manual or automatic. In both cases, the UE scans all possible RF channels that it supports based on its capability. During the search, and specifically for LTE, all detected PLMNs with RSRP ≥ -110 dBm are reported to the UE's NAS as high-quality PLMNs without an associated RSRP value.[1] PLMNs with an RSRP < -110 dBm are reported along with an associated RSRP value. The order in which the UE searches for PLMNs (either intra- or inter-band) is not defined but can be optimized based on information that the UE already has such as the last frequency on which it successfully camped. In both automatic and manual modes, the search for PLMNs can be superseded by the NAS at any time.

When the UE is in automatic mode and multiple PLMNs are detected, the order in which selection is performed is governed by a set of rules such that the last registered PLMN is tried first, then the home PLMN, followed by a user-defined and operator-defined list of PLMNs in priority order, followed by high-

[1]Reference signal received power (RSRP) is an important UE physical layer measurement and is the linear average (in watts) of the downlink cell-specific reference signals across the channel bandwidth. Since the cell-specific reference signals exist only for one symbol at a time, the measurement is made only on those resource elements that contain cell-specific reference signals. It is not mandated for the UE to measure every reference signal on the relevant sub-carriers. Instead, accuracy requirements have to be met. There are requirements for both absolute and relative RSRP. The absolute requirements range from ± 6 to ± 11 dB depending on the noise level and environmental conditions. Measuring the difference in RSRP between two cells on the same frequency (intra-frequency measurement) is a more accurate operation for which the requirements vary from ± 2 to ± 3 dB. The requirements are relaxed to ± 6 dB when the cells are on different frequencies (inter-frequency measurement).

Knowledge of absolute RSRP provides the UE with essential information about the strength of cells from which path loss can be calculated and used in the algorithms for determining the optimum power settings for operating the network. Reference signal received power is used both in idle and connected modes. The relative RSRP is used as a parameter in multi-cell scenarios [34].

Although RSRP is an important measure, on its own it gives no indication of signal quality. The reference signal received quality (RSRQ) provides this measure and is defined as the ratio of RSRP to the E-UTRA carrier received signal strength indicator (RSSI). The RSSI parameter represents the entire received power including the desired signal power from the serving cell as well as all co-channel power and other sources of noise. Measuring RSRQ becomes particularly important near the cell-edge when decisions need to be made, regardless of absolute RSRP, to perform a handover to another cell. Reference signal received quality is used only during connected states. Intra- and inter-frequency absolute RSRQ accuracy varies from ± 2.5 to ± 4 dB, which is similar to the inter-frequency relative RSRQ accuracy of ± 3 to ± 4 dB [34].

Samsung Ex. 1005, Page 189 of 1147

quality PLMNs in random order, followed by other PLMNs in order of decreasing RSRP are tried. If a UE is camped on a PLMN that is not the highest priority, it will periodically search for higher priority PLMNs and report the results to the UE's NAS. The interval between searches is stored on the USIM and set by the service provider to be between 6 min and 8 hour, in increments of 6 min [33].

Cell search is a process through which the UE acquires the RF carrier frequency, cyclic prefix length (or OFDM symbol timing), subframe and frame timing, and the physical cell identifier. The E-UTRA uses a hierarchical cell search mechanism which is conducted by acquiring the primary synchronization signals followed by acquiring the secondary synchronization signals. In order to support cell search over scalable transmission bandwidth, E-UTRA requires the UEs to support a minimum bandwidth of 20 MHz in the downlink. The cells in E-UTRA are classified as follows [5]:

- Serving cell: The cell on which the UE is currently camped.
- Suitable cell: A cell that the UE may camp on to obtain normal service, which is a cell whose measured cell attributes satisfy the cell selection criteria; the cell is either part of the selected PLMN, registered or an equivalent PLMN; the cell is not barred or reserved, and the cell is not part of a tracking area that is in the list of forbidden tracking areas for roaming. If more than one PLMN ID is broadcast in the cell, the cell is considered to be part of all tracking areas with tracking area identities (TAIs) constructed from the PLMN IDs and the TAC broadcast in the cell.
- Acceptable cell: A cell whose measured cell attributes satisfy the cell selection criteria and is not barred. The UE may camp on this cell to obtain limited service to originate emergency calls and receive earthquake and tsunami warning system (ETWS) notifications.
- Barred cell: A cell that the UE is not allowed to camp on.
- Restricted cell: A cell on which camping is allowed, but access attempts are disallowed for UEs whose access classes are indicated as barred.
- Reserved cell: A cell on which camping is not allowed, except for particular UEs, if that is indicated in the system information.
- Closed-subscriber group (CSG) cell: A CSG cell is accessible by the members of the CSG for that CSG ID. A cell in which only members of the CSG can obtain normal service. Depending on local regulation, the CSG cell can provide emergency bearer services also to subscribers who are not members of the CSG [3].

Following successful completion of the cell search, the UE acquires PBCH and obtains the MIB, which carries the overall downlink transmission bandwidth, PHICH configuration, and the system frame number (SFN). The UE then decodes PDCCH in order to locate the PDSCH resources carrying SIBs which are common radio resource and mobility configurations for all UEs in the cell (see Figure 4.2).

Synchronization signals are transmitted in every radio frame. The synchronization signals are used for time and frequency synchronization during initialization. As part of the system acquisition process, the UE synchronizes to OFDM symbol, slot, subframe, half-frame, and radio frame (in that order) using synchronization signals. Two synchronization signals are defined for system acquisition in E-UTRA: (1) Primary synchronization signal is used to obtain slot, subframe, and half-frame boundary. It also provides cell ID within the cell ID group. (2) Secondary synchronization signal is used to obtain the radio frame boundary. It also enables the UE to determine the cell ID group, which may range from 0 to 167. The synchronization signals are transmitted in the 0th and 5th subframes of every radio frame. In the frequency domain, the synchronization signals occupy the center 1.08 MHz (equivalent to six resource blocks) of the radio channel.

In the time domain, primary and secondary synchronization signals are transmitted in the 0th and 10th slot (1st slot within 0th and 5th subframe) of every radio frame. Within these slots, the primary and secondary synchronization signals are transmitted in the last two symbols. The primary synchronization signal is transmitted in the last symbol of 0th and 10th slot. The secondary synchronization signal is transmitted in one before the last symbol of 0th and 10th slot. The PCBH is mapped to the first four symbols in the 1st slot of the frame (just after primary and secondary synchronization signals). In the frequency domain, the center 1.08 MHz of the radio channel is reserved for primary and secondary synchronization signals. This is equivalent to 6 resource blocks or 72 subcarriers. However, the primary and secondary synchronization signals only require 62 subcarriers around the DC subcarrier. The remaining subcarriers (10 subcarriers) in these resource blocks are not used for transmission. They provide a guard band between primary and secondary synchronization signal transmission and other data transmissions. There are 504 unique physical-layer cell IDs. The physical-layer cell IDs are divided into 168 unique physical-layer cell ID groups, each group contains three unique IDs. The grouping is done such that each physical-layer cell ID is part of one and only one physical-layer cell ID group. A physical-layer cell ID $N_{\mathrm{ID}}^{\mathrm{cell}} = 3N_{\mathrm{ID}}^{(1)} + N_{\mathrm{ID}}^{(2)}$ is thus uniquely defined by a number $N_{\mathrm{ID}}^{(1)}$ in the range of 0−167, representing the physical-layer cell ID group, and a number $N_{\mathrm{ID}}^{(2)}$ in the range of 0−2, denoting the physical-layer ID within the physical-layer cell ID group [20−22,33].

The essential system information which allows the other channels in the cell to be configured and operated is carried by PBCH. Therefore, the reliability and coverage of PBCH is crucial to the successful operation of E-UTRA systems. The SI is divided into two categories: (1) the MIB, which consists of a limited number of the most frequently transmitted parameters essential for initial access to the cell which is carried in PBCH; and (2) the other SIBs which are multiplexed with unicast data and are transmitted in PDSCH.

The detectability of the PCBH without UE's prior knowledge of the system bandwidth is achieved by mapping PBCH to the center of the band using 72

subcarriers, which corresponds to the minimum system bandwidth of six resource blocks, regardless of the actual system bandwidth. The UE will have to identify the system center frequency from the synchronization signals. Minimizing the impact on total overhead is achieved by deliberately keeping the amount of information carried by PBCH to a minimum, since achieving stringent coverage requirements for a large quantity of data would result in a high system overhead. The size of MIB is 24 bits and is transmitted every 40 ms. In order to achieve reliable reception, time diversity, forward error correction, and antenna diversity mechanisms are used to transmit PBCH. Time diversity is introduced by spreading the transmission of PBCH over a period of 40 ms with three retransmissions at 10 ms intervals. This significantly reduces the likelihood of missing/misdetection of MIB due to multipath fading, interference, and noise in the radio channel, even when the mobile terminal is moving at high speeds.

The forward error correction coding for PBCH utilizes a tail-biting convolutional coder, considering that the number of information bits to be encoded is small and turbo codes are typically not suitable for small payloads. The basic code rate is 1/3, after which a high repetition factor is applied to the systematic and parity bits such that each MIB is coded at a very low code rate (1/48 over a 40 ms period) to provide strong error protection. Antenna diversity may be utilized at both the eNB and the UE. The UE performance requirements specified for E-UTRA assume that all UEs can achieve a level of decoding performance proportional to dual-antenna receive diversity although it is understood that in low-frequency deployments, e.g., below 1 GHz, the advantage obtained from receive diversity is reduced due to the correspondingly higher correlation between the antennas. This enables wider or more reliable cell coverage to be achieved with fewer cell sites.

Transmit antenna diversity may also be employed at the eNB to further improve coverage. Depending on eNB implementation, it may transmit PBCH with two or four transmit antennas using space-frequency block codes (SFBC). The set of resource elements designated to PBCH transmission is independent of the number of transmit antennas at the eNB, thereby any resource elements which may be used for cell-specific reference signals transmission are avoided by PBCH, irrespective of the actual number of transmit antenna ports used by the eNB. The number of transmit-antennas used by the eNB is blindly detected by the UE by decoding each SFBC scheme corresponding to different number of transmit antennas (i.e., one, two, or four). The detection of the number of transmit antennas is further facilitated by masking the cyclic redundancy check (CRC) bits of each MIB with a unique masking sequence representing the number of transmit antenna ports.

Achieving low latency and minimal impact on UE power consumption are assisted through use of a low code rate with repetition, wherein the set of coded bits are divided into four subsets, each of which is self-decodable. Each subset of the coded bits is transmitted in a different radio frame during the 40 ms transmission period. This means that if the SINR is sufficiently high to allow

the UE to successfully decode the MIB in less than four radio frames, then the UE does not need to receive the other parts of the PBCH transmission in the remainder of the 40 ms period. On the other hand, if the SINR is low, the UE can take advantage of soft-combing of other parts of the MIB transmission until the MIB is successfully decoded. The UE does not know in advance the timing of the 40 ms transmission interval for each MIB on PBCH; therefore, this information is determined implicitly from the scrambling and bit positions, which are reinitialized every 40 ms. The UE can initially determine the 40 ms timing by performing four separate decodings of the PBCH using each of the four possible phases of the PBCH scrambling code and checking the CRC for each decoding.

When a UE initially attempts to select a cell by acquiring PBCH, different approaches may be taken to perform the necessary blind decoding. A simple approach is to perform the decoding using a soft combing of the PBCH instances over four radio frames, advancing a 40 ms sliding window one radio frame at a time until the sliding window aligns with the 40 ms period of the PBCH and the decoding succeeds. However, this would result in a 40–70 ms delay before the PBCH can be decoded. A faster approach would be to attempt to decode the PBCH from the first radio frame, which should be possible provided that the SINR is sufficiently high. If the decoding fails for all four possible scrambling sequences, the PBCH from the first frame could be soft-combined with the next PBCH instances. It is evident that the latter approach may be much faster at the expense of slightly more complexity.

We learned so far that initial frequency and timing acquisition are based on the synchronization signals (Figure 4.3). A minimum bandwidth of 1.08 MHz (six resource blocks) is used for the synchronization and broadcast channels irrespective of the actual system bandwidth. The UE acquires a coarse frequency synchronization; OFDM symbol, slot, and subframe timing; and the index to a cell in the cell ID group through successful detection and processing of the primary synchronization signals. The UE further acquires the half-frame timing; radio frame timing; cell ID group; and FDD/TDD mode information via the secondary synchronization signals. Note that there is no eNB antenna configuration information on the secondary synchronization signals. The eNB antenna configuration is conveyed in PBCH by applying different CRC masks. The cell search procedure is complete when PBCH CRC check passes.

In E-UTRA, the MIB, SIB1, and SIB2 broadcast messages are transmitted for any cell in the network in such a way that the location (subframe) where the SIB is transmitted should not be the same subframe where another SIB is transmitted. As shown in Figure 4.4, the SIB scheduling follows these principles: (1) MIB is transmitted at fixed cycles every four frames starting from SFN 0; (2) SIB1 is also transmitted at fixed cycles every eight frames starting from SFN 0; (3) all other SIBs are transmitted at the cycles between 80 and 5120 ms that are specified by SIB scheduling information elements in SIB1. The SIB1 uses a fixed schedule with a periodicity of 80 ms with retransmissions within 80 ms intervals.

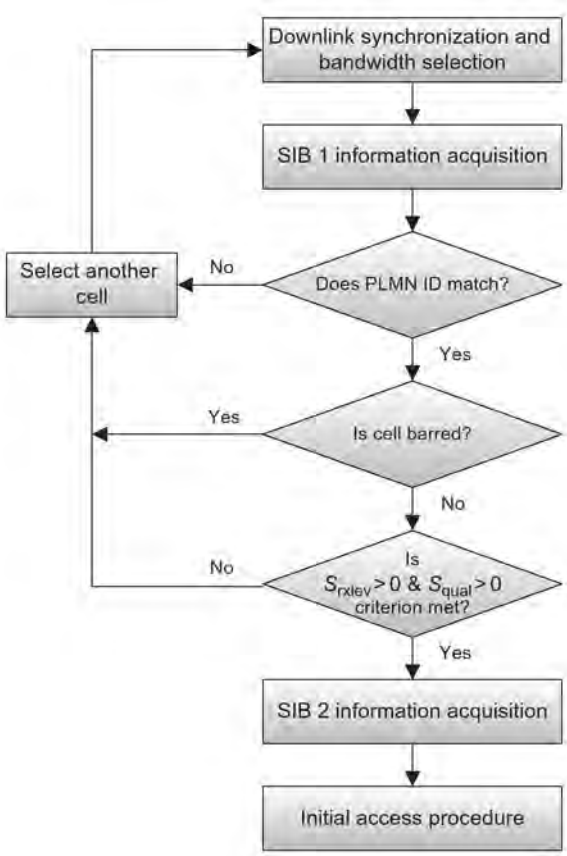


FIGURE 4.3

Cell selection summary [20–22,33].

The first transmission of SIB1 is scheduled in the 5th subframe of radio frames for which SFN mod8 = 0 and retransmissions are scheduled in the 5th subframe of all other radio frames for which SFN mod2 = 0. This means that even though SIB1 periodicity is 80 ms, different redundancy versions of the SIB1 are transmitted every 20 ms. It means that at layer-3 level, the SIB1 is transmitted every 80 ms, but at layer-1 level, it is seen every 20 ms [9].

The transmission periods for other SIBs are determined by the *schedulingInfoList* information element in SIB1 as shown in the following example. In this example, SIB2 and SIB3 are transmitted every 160 and 320 ms, respectively:

```
+-schedulingInfoList::= SEQUENCE OF SIZE(1..maxSI-Message[32]) [2]
|+-SchedulingInfo::= SEQUENCE
||+-si-Periodicity::= ENUMERATED [rf16]
||+-sib-MappingInfo::= SEQUENCE OF SIZE(0..maxSIB-1[31]) [0]
|+-SchedulingInfo::= SEQUENCE
```

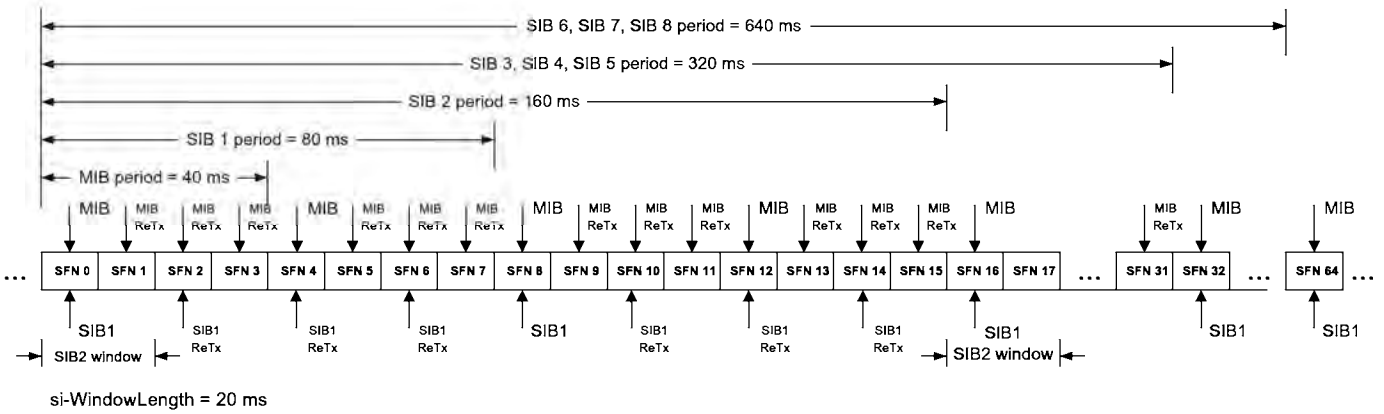


FIGURE 4.4

System information transmission timing (example).

```
| + -si-Periodicity:: = ENUMERATED [rf32]
| + -sib-MappingInfo:: = SEQUENCE OF SIZE(0..maxSIB-1[31]) [1]
| + -SIB-Type:: = ENUMERATED [sibType3]
 + -tdd-Config:: = SEQUENCE OPTIONAL:Omit
 + -si-WindowLength:: = ENUMERATED [Ms20]
```

You would also note that the *sib-MappingInfo* information element in the first node is not specified, but the first entity of *schedulingInfoList* should always be for SIB2 as specified in 3GPP TS 36.331 [9]. There is no mapping information of SIB2 since it is always present in the first SI message listed in *schedulingInfoList* list.

Understanding the overall transmission cycle and location of SIBs is related to interpretation of the *si-WindowLength* parameter, which specifies that the corresponding SIB should be transmitted somewhere within the window length starting at the SFN specified by *si-Periodicity*. However, this parameter does not specify the exact subframe number for the SIB transmission. The subframe for a specific SIB transmission is determined by an algorithm defined in 3GPP TS 36.331 [9], and described in the following. When acquiring an SI message, the UE is required to determine the start of the *SI-window* for the corresponding SI message by calculating the entry number n which corresponds to the order of entry in the list of SI messages configured by the *schedulingInfoList* in SIB1. It then calculates the integer-valued number $x = (n-1)w$, where w is given by *si-WindowLength*. The SI-window starts at subframe number $x \bmod 10$ in the radio frame for which SFN $\bmod T_{\text{SIB}} = \lfloor x/10 \rfloor$, where T_{SIB} is given by *si-Periodicity* of the corresponding SI message. Note that E-UTRAN should configure an SI-window of 1 ms only if all SIs are scheduled before the 5th subframe in radio frames for which SFN mod2 = 0. Therefore, the UE must decode the downlink shared channel using the SI-RNTI from the start of the SI-window and continue until the end of the SI-window whose absolute length in time is given by *si-WindowLength*, or until the SI message was correctly received, with the exception of the following subframes: (1) the 5th subframe in radio frames for which SFN mod2 = 0; (2) any MBSFN subframes; and (3) any uplink subframes in TDD. If the SI message was not received by the end of the SI-window, the UE repeats this process at the next SI-window occasion for the corresponding SI message [9,21,32].

The UE uses one of the following two cell selection procedures [5]:

1. Initial cell selection: This procedure requires no prior knowledge of E-UTRA RF carriers. The UE scans all RF channels in the E-UTRA bands according to its capabilities to find a suitable cell. On each carrier frequency, the UE needs only to search for the strongest cell. Once a suitable cell is found, that cell is selected.
2. Stored information cell selection: This procedure requires stored information of carrier frequencies and also information on cell parameters from previously received measurement control information elements or from previously detected cells. Once a suitable cell is found, the UE selects that cell. If no suitable cell is found, the initial cell selection procedure is enacted.

Samsung Ex. 1005, Page 196 of 1147

The SIB1 specifies the criteria for determining whether a cell is suitable for selection [9]. A cell is considered suitable when $S_{rxlev} > 0$ & $S_{qual} > 0$ where for LTE, S_{rxlev} denotes the cell selection receive level value in dB and S_{qual} is the cell selection quality value in dB. The detailed cell selection and reselection criteria are described in Chapter 5. In practice, a UE receives signals from several cells belonging to different network operators. The UE only knows, after reading the SIB1 content (PLMN ID), if this cell belongs to the operator's network to which it is subscribed. Therefore, the UE will initially look for the strongest cell per carrier, then for the PLMN ID by decoding the SIB1 content to decide whether that is a suitable network. It will then compute the $S_{rxlev} > 0$ & $S_{qual} > 0$ criterion and decide whether the cell is suitable.

As an example, assume that the UE belongs to network operator A and there are two other operators which are operating LTE networks but at different frequencies. The UE receives signals of all base stations but at different power levels. Based on the above definition, the UE will select the strongest cell that may belong to network operator C; however, it will reject the cell after decoding the SIB1 content and observing that the PLMN ID saved on the USIM card does not match to that of the cell. The UE will now proceed to the next strongest signal, which may belong to network operator B, and when the PLMN ID does not correspond to the UE's stored information, it will continue with the cell belonging to network operator A. Once the PLMN ID matches and PLMN is selected, the UE will continue to use the information in SIB1 and SIB2 to evaluate the cell selection criteria based on signals received from eNBs belonging to network A. If the parameters transferred and belonging to eNB_1 do not fulfill $S_{rxlev} > 0$ & $S_{qual} > 0$ criterion, the UE will continue with demodulating and decoding the information provided by eNB_2 and if the $S_{rxlev} > 0$ & $S_{qual} > 0$ criterion is fulfilled, the UE will camp on this cell (see Figure 4.3).

4.3 EPS connection and mobility management

The states of a UE with respect to mobility and connection establishment are described by the following NAS states: the EPS mobility management (EMM) state (EMM-DEREGISTERED or EMM-REGISTERED) reflects whether the UE is registered at the MME and the EPS connection management (ECM) state (ECM-IDLE or ECM-CONNECTED) reflects the connectivity of the UE with the EPC. The AS and NAS states and state transitions are illustrated in Figure 4.5. The transition from ECM-IDLE to ECM-CONNECTED not only involves establishment of the RRC connection, but also includes establishment of S1 connection. The RRC connection establishment is initiated by the NAS and is completed prior to S1 connection establishment, which means that connectivity in RRC_CONNECTED is initially limited to the exchange of control information between the UE and eNB. The UEs are typically moved to ECM-CONNECTED

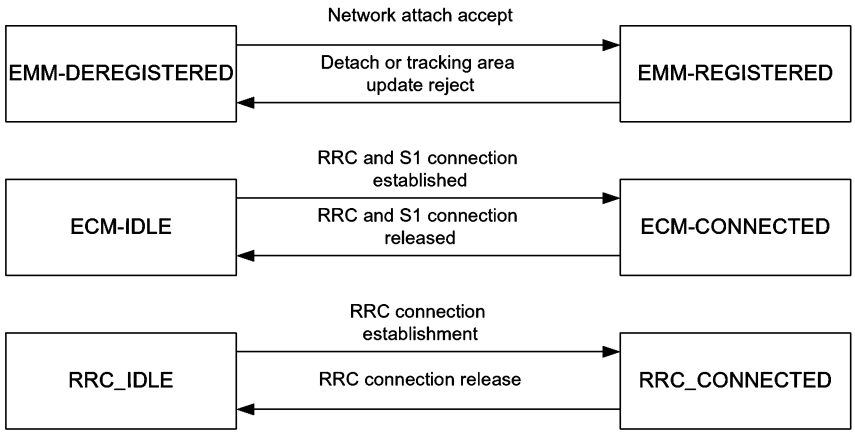


FIGURE 4.5

AS and NAS states and state transition [21].

when becoming active. It should be noted that in LTE, the transition from ECM-IDLE to ECM-CONNECTED is performed within 100 ms and the UEs engaged in intermittent data transfer do not need to stay in ECM-CONNECTED, if the ongoing services can tolerate such transfer delays. RRC connection release is initiated by the eNB following release of the S1 connection between the eNB and the core network [20].

4.3.1 EPS mobility management

The EMM states describe the mobility management states that result from the mobility management procedures, e.g., attach and tracking area update procedures. The EMM has two states: EMM-REGISTERED and EMM-DEREGISTERED as shown in Figure 4.5. In EMM-DEREGISTERED state, the EMM context does not have any valid location or routing information for the UE. Since the location of the UE is not known, the UE is not reachable by the MME. In order to avoid performing an authentication and key agreement (AKA)[2] security procedure in every network attach attempt by the UE, the MME and the UE may store some UE context in this state. The MME may perform an implicit detach any time after the implicit detach timer expires. The state is changed to EMM-DEREGISTERED in the MME after performing the implicit detach. If all bearers belonging to a UE are released (e.g., after a handover from E-UTRAN to

[2]AKA is a mechanism which performs authentication and session key distribution in UMTS networks. AKA is a challenge-response based mechanism that uses symmetric cryptography. AKA is typically run in a UMTS IP multimedia services identity module (ISIM), which resides on a smart card device that also provides tamper resistant storage of shared secrets. AKA is defined in IETF RFC 3310.

non-3GPP access [11]), the MME changes the mobility management state of the UE to EMM-DEREGISTERED. This ensures that the UE performs a tracking area update when it reselects E-UTRAN [4].

In EMM-REGISTERED state, the UE has registered with the MME following an attach procedure. When the UE transitions to this state, a UE context and a default EPS bearer have been established and the UE can receive services that require registration with the EPC. The location of the UE is known by the MME at the tracking area level. In the EMM-REGISTERED state, an idle UE is required to perform a tracking area update with the network. The detach procedure or a tracking area update reject changes the EMM state from EMM-REGISTERED to EMM-DEREGISTERED. In the EMM-REGISTERED state, the UE always has at least one active PDN connection (IP connection) and EPS security context [4].

In general, the ECM and EMM states are independent of each other. Transition from EMM-REGISTERED to EMM-DEREGISTERED can occur regardless of the ECM state, e.g., by explicit detach signaling in ECM-CONNECTED or by implicit detach locally in the MME during ECM-IDLE. However, there are some dependencies between the two states, e.g., to transition from EMM-DEREGISTERED to EMM-REGISTERED the UE has to be in the ECM-CONNECTED state [20].

4.3.2 EPS connection management

The ECM states describe the signaling connectivity between the UE and the EPC. In EMM-REGISTERED state, there are two EPS connection states as follows: ECM-IDLE and ECM-CONNECTED as shown in Figure 4.5. In ECM-IDLE, a UE performs cell selection or reselection. There is no NAS signaling connection between the UE and the network. The UE does not have a context in the E-UTRAN and has no S1-MME and S1-U connection with the network. In EMM-REGISTERED and ECM-IDLE state, the UE performs a tracking area update, if the current TA is not in the list of TAs that the UE has received from the network in order to maintain the registration and enable the MME to page the UE. It performs a periodic tracking area update procedure to notify the EPC that the UE is available. The UE performs a tracking area update, if the RRC connection was released with release cause "*load balancing TAU required*" or performs a tracking area update if there are any changes in the UE's core network capability information or the UE-specific DRX parameters. The UE performs a tracking area update when it manually selects a CSG femto-cell, and the CSG ID[3] of that cell is absent from both the UE's allowed CSG list and the UE's operator CSG list. The UE answers to paging from the MME by performing a service request procedure and performs the service request procedure in order to establish the radio bearers when uplink user data is available for transmission.

[3]A CSG ID is a unique identifier within the scope of one PLMN defined in 3GPP TS 23.003 [1] which identifies a CSG in the PLMN associated with a cell or group of cells to which access is restricted to members of the CSG.

The UE and the MME enter the ECM-CONNECTED state when the signaling connection is established between the UE and the MME. Initial NAS messages that initiate a transition from ECM-IDLE to ECM-CONNECTED state are *Attach Request*, *Tracking Area Update Request*, *Service Request*, or *Detach Request*. When the UE is in ECM IDLE state, the UE and the network may be unsynchronized, i.e., the UE and the network may have different sets of established EPS bearers. When the UE and the MME transition to the ECM-CONNECTED state, the set of EPS bearers are synchronized between the UE and the network.

When signaling connection is established between the UE and the network, the UE and the MME make transition to the ECM-CONNECTED state. In the ECM-CONNECTED state, the UE location is known in MME at cell level. The UE can perform handover and tracking area update procedures. The UE has a signaling connection with the MME. A service request upon availability of uplink data or paging message upon availability of downlink data change the UE state from ECM-IDLE to ECM-CONNECTED. The UE-assisted measurements are required for handover, cell selection, and reselection. The signaling connection is made up of two parts: an RRC connection and an S1-MME connection. The UE transitions to the ECM-IDLE state when its signaling connection to the MME has been released or broken. This release or failure is explicitly indicated by the eNB to the UE or detected by the UE. The S1 release procedure changes the state at both UE and MME from ECM-CONNECTED to ECM-IDLE. The UE may not receive the indication for the S1 release due to radio link error or out of coverage causes. In this case, there can be temporary mismatch between the ECM state in the UE and the ECM state in the MME [4,5].

4.4 UE states

During normal operation, an LTE UE can be in one of the two (steady-state) RRC states: RRC_IDLE and RRC_CONNECTED, as shown in Figure 4.6. The emphasis on the steady-state descriptor is due to the fact that when a UE is powered on or upon radio link failure, it goes to a transient state with no attachment to a PLMN or a cell, thus with no connection to the network. The UE starts scanning and selecting a PLMN and then proceeds to camp on a cell (cell selection/reselection) through downlink synchronization and system information acquisition. In this case, the UE has virtually moved to the RRC_IDLE state since it still has no RRC connection, and thus performs uplink synchronization and RRC connection setup procedures. In RRC_IDLE, the UE monitors a paging channel and/or SIB1 content to detect system information change and other system notifications. The UE further monitors control channels associated with the shared data channel to determine whether data is scheduled for it. The RRC_CONNECTED is a state where the UE transmits or receives unicast control/data in the uplink or downlink and may be configured with a customized DRX pattern. The UE monitors control channels associated with the shared data channel to determine if data is scheduled for it, provides channel quality and feedback information, performs

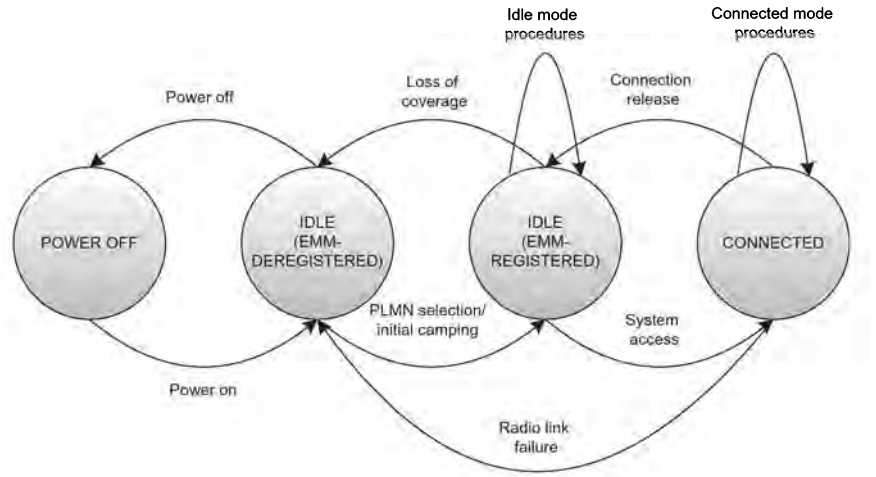


FIGURE 4.6

E-UTRA UE states and state transitions [20–22,32].

neighboring cell measurements and measurement reporting, and acquires the system information. One or multiple IP addresses are assigned to the UE as well as a temporary identifier called C-RNTI that is used for signaling purposes between the UE and the network.

Prior to transition to RRC steady-states, the UE uses the information stored on its USIM and the local memory for system selection and camping. Upon initial power-up, the UE uses the above static information to initiate system selection. In a practical scenario, the IMSI on the USIM may be used to determine the home PLMN code and may serve as the basis for PLMN selection. The USIM may have optionally an equivalent home PLMN list and priorities among them to determine the PLMN selection. The USIM may also provide the allowed RAT for each PLMN where there can be multiple RATs for a given PLMN. The UE considers the supported RAT and frequency band to trigger system selection. The system selection procedures are divided between NAS and AS. Over the air, signaling support provided by the AS assists NAS in camping on a PLMN. The selected PLMN is known as the registered PLMN and is stored on the USIM for future system selections.

The idle mode procedures depend on the UE's camping status. As shown in Figure 4.6, there are two idle states that a UE may transition to: (1) Idle with EMM-DEREGISTERED; and (2) Idle with EMM-REGISTERED. When the UE is in the Idle EMM-DEREGISTERED state, no suitable cell on any PLMN or RAT has been selected; therefore, this state triggers the cell search procedure in order to detect a cell that is suitable to camp on. When the UE is in the Idle EMM-REGISTERED state, it has acquired a suitable cell belonging to a non-barred/unrestricted PLMN. While in this state, the UE performs cell reselection on the selected PLMN; searches for high-priority PLMNs and a suitable cell, if another

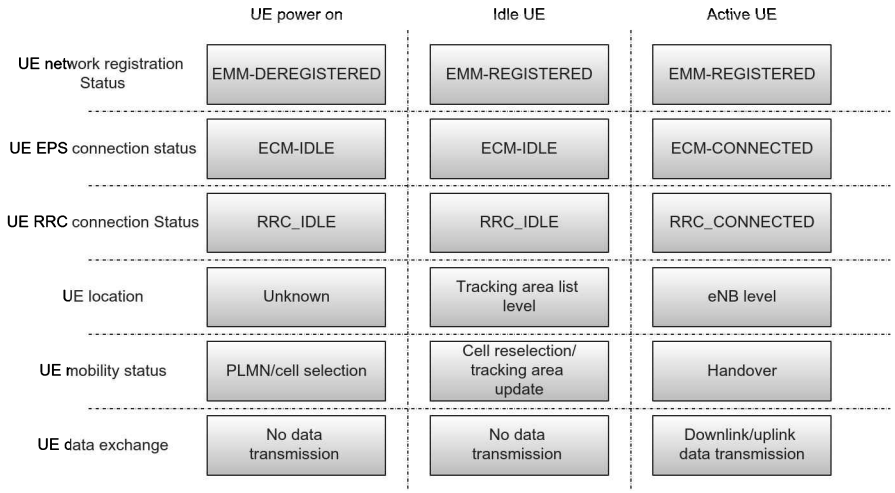


FIGURE 4.7

Summary of the UE and EPS states and the UE status in each state.

PLMN is selected by NAS; monitors paging messages; updates SI; initiates access procedures to transition to connected mode; and performs location update.

In the state diagram shown in Figure 4.6, the connected state refers to the EMM-REGISTERED + ECM-CONNECTED + RRC_CONNECTED state, the idle/EMM-REGISTERED refers to the EMM-REGISTERED + ECM-IDLE + RRC_IDLE state, and the idle/EMM-DEREGISTERED refers to the EMM-DEREGISTERED + ECM-IDLE + RRC_IDLE state. It must be noted that the Idle/EMM-DEREGISTERED state is a transient state, whereas the Idle/EMM-REGISTERED and connected states are steady-state states. Figure 4.7 provides a more thorough perspective into UE status in each of the UE and EPS states. As shown in the figure, when the UE is powered on, it must find a suitable PLMN, select a cell and register with the network. An idle UE might have just terminated its data exchange with the network, releasing RRC and ECM signaling connections. An active UE has attached to the EPS core network and established a default EPS bearer. In this state, the UE is EMM-REGISTERED + ECM-CONNECTED + RRC_CONNECTED, and a UE context is created and stored in the UE, eNB, and MME. In addition, the figure also shows the mobility and location status and information in each state of the UE and EPS mobility and connection management.

4.4.1 Idle mode procedures

The idle mode procedures can be divided into four main tasks, namely PLMN selection; cell selection and reselection; location registration; and support for manual femto-CSG selection. The relationship between these processes is

illustrated in Figure 4.8. A UE supporting CSG selection selects a CSG femtocell either automatically based on the list of allowed CSG IDs, or manually based on user selection of CSG within the list of available CSGs.

When a UE is switched on, a PLMN is selected by NAS, where associated RAT(s) may be set for the selected PLMN. The NAS provides a list of equivalent PLMNs (if any) that the AS may use for cell selection and cell reselection. During cell selection, the UE searches for a suitable cell belonging to the selected PLMN and selects a suitable cell to gain network access. This process is known as camping. The UE must then proceed with registration with the network by means of a NAS registration procedure in the tracking area of the chosen cell, and as the outcome of a successful location registration, the selected PLMN becomes the registered PLMN. If the UE finds a more suitable cell, according to the cell reselection criteria, it reselects that cell and camps on it. If the new cell does not belong to at least one tracking area to which the UE is registered, location registration is repeated. The UE searches for higher priority PLMNs at regular time intervals and searches for a suitable cell if another PLMN has been selected by NAS. Camping on a cell is necessary in order to gain access to network services. Three levels of services may be provisioned for a UE: (1) limited service (emergency calls, ETWS, and commercial mobile alert service (CMAS) on an acceptable cell); (2) normal service (for public use on a suitable cell); and (3) operator service (for operators only on a reserved cell) [5].

The search for available femto-CSGs may be triggered by the UE's NAS to support manual femto-CSG selection. If the UE experiences loss of coverage in

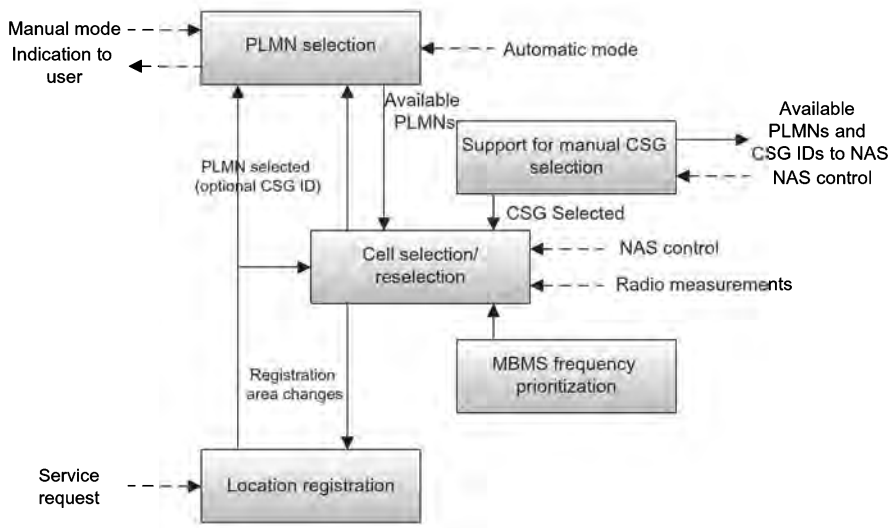


FIGURE 4.8

Overview of idle mode procedures [5].

the registered PLMN, either a new PLMN is selected automatically in automatic mode or an indication of which PLMNs are available is given to the user, so that a manual selection can be made. The purpose of camping on a cell in the idle mode is to enable the UE to receive SI from the network. Furthermore, if the network receives a call or downlink data for a registered UE, it knows the set of tracking areas in which the UE is camped, thus it can then send a paging message for the UE on the control channels of all cells in the set of tracking areas to which the UE is associated. The UE will then receive the paging message because it is tuned to the control channel of a cell in one of the registered tracking areas and the UE can respond on that control channel. It further enables the UE to receive MBMS services while in the idle mode. The tracking area update procedure and signaling is depicted in Figure 4.9. The process starts with RRC connection establishment and is followed by the UE's request for tracking area update,

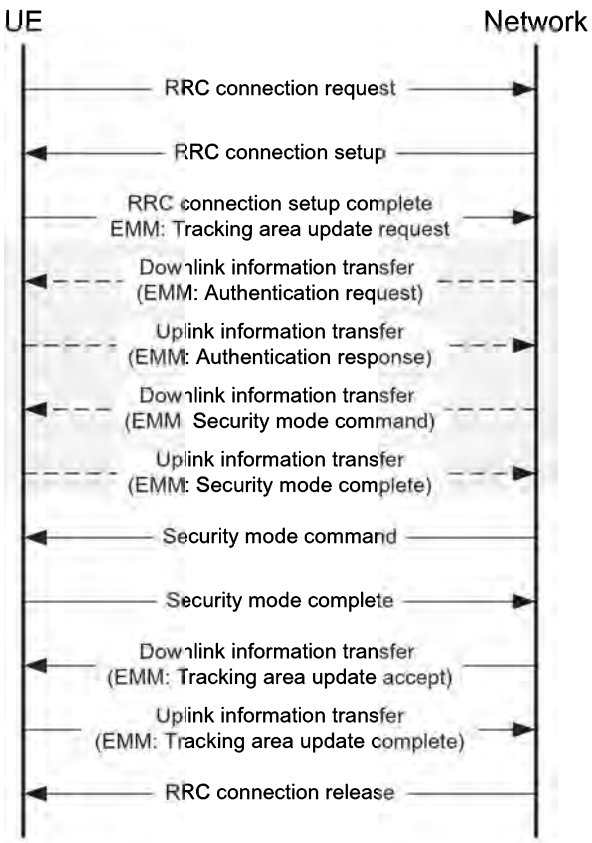


FIGURE 4.9

Tracking area update procedure and signaling [5,25].

authentication and security commands, and finally release of the RRC connection upon successful update of tracking area information. If the UE is unable to find a suitable cell to camp on, or if the location registration fails, it attempts to camp on any cell irrespective of the PLMN ID to receive limited services.

The procedures performed by the UE while in the idle mode, as shown in Figure 4.8, are divided into AS and NAS functions. Table 4.1 provides the functional division between the UE's NAS and AS functions in the idle mode.

4.4.1.1 Tracking area update

After the UE successfully attaches to the network, the UE can move around in the assigned tracking area. If it detects a different tracking area, it needs to inform the network of the new tracking area. Similarly, the network can request the UE to update its tracking area periodically, which helps the network to locate the UE faster whenever a mobile-terminated call is received by the network for the UE since it can just page the UE in the last reported tracking area. During the tracking area update procedure, the MME may initiate an authentication procedure and set up a security context. The tracking area update procedure and signaling is illustrated in Figure 4.9.

4.4.1.2 Paging

The paging procedure is used by the network to request establishment of NAS signaling connection to the UE. If there is an IP packet that arrives for a UE from an external network to the P-GW and there is no existing dedicated bearer for the UE, it will forward the IP packet to the S-GW on the default bearer. Once the packet has reached the S-GW on the default bearer, the S-GW detects the need to create a dedicated bearer and sends the *Downlink Data Notification* message to the MME in order to page the UE and to create the dedicated bearers. At this time, the MME has to ensure that the UE establishes an RRC connection, so the MME sends a *Paging Request* message to all eNBs associated with the last known tracking area in which the UE was located [25].

In other words, when there is a need to deliver downlink data to an ECM-IDLE UE, the MME sends a paging message to all eNBs in its current tracking area, and the eNBs page the UE over the radio interface. Upon receipt of a paging message, the UE performs a service request procedure, which results in transition to ECM-CONNECTED state. The UE-related information is thereby created in the E-UTRAN, and the radio bearers are reestablished. The MME is responsible for reestablishment of the radio bearers and updating the UE context in the eNB. To accelerate the idle-to-active transition and bearer establishment, EPS supports concatenation of the NAS and AS procedures for bearer activation. Some relationship between the NAS and AS protocols is employed to allow procedures to run simultaneously rather than sequentially. For example, the bearer establishment procedure can be executed by the network without waiting for the completion of the security procedure. The security functions are the responsibility of the MME for both signaling and user data. When a UE attaches to the network, a mutual authentication of the UE and the network is performed between the UE and the

Table 4.1 AS and NAS Functions in Idle Mode [5]

Idle Mode Task	UE's NAS Functions	UE's AS Functions
PLMN selection	Maintain a list of PLMNs in priority order Select a PLMN using automatic/manual mode and request AS to select a cell belonging to this PLMN Evaluate reports of available PLMNs from AS for PLMN selection Maintain a list of equivalent PLMN IDs	Search for available PLMNs Search in this set of RAT(s) and other RAT(s) under the selected PLMN Perform measurements for PLMN selection Synchronize to a broadcast channel to identify scanned PLMNs Report available PLMNs with associated RATs to UE's NAS upon request from NAS or autonomously
Cell selection	Control cell selection by indicating RATs associated with the selected PLMN to be used initially in the search of a cell in the cell selection process NAS maintains lists of forbidden registration areas and a list of CSG IDs and their associated PLMN ID on which the UE is allowed (CSG whitelist) and provides these lists to AS	Perform measurements for cell selection Detect and synchronize to a broadcast channel Receive and process broadcast information Forward NAS system information to UE's NAS Search for a suitable cell. The cells broadcast one or more PLMN IDs in their system information If associated RATs are set for the selected PLMN, perform the search among these RATs and other RATs for that PLMN The cell is camped on when it is found
Cell reselection	Control cell reselection by, for example, maintaining lists of forbidden registration areas Maintain a list of equivalent PLMN IDs and provide the list to AS Maintain a list of forbidden registration areas and provide the list to AS Maintain a list of CSG IDs and their associated PLMN ID on which the UE is allowed (CSG whitelist) to camp and provide the list to AS	Perform measurements to support cell reselection Detect and synchronize to a broadcast channel Receive and process broadcast information and forward NAS system information to UE's NAS Change cell when a more suitable cell is found

(*Continued*)

Table 4.1 (Continued)

Idle Mode Task	UE's NAS Functions	UE's AS Functions
Location registration	Register the UE as active after power on	Report registration area information to NAS
	Register the UE's presence in a registration area, for instance regularly or when entering a new tracking area	
	Maintain lists of forbidden registration areas	
	Deregister UE when shutting down	
Support for manual femto-CSG selection	Provide request to search for available femto-CSGs	Search for cells with the specified CSG ID
	Evaluate reports of available femto-CSGs from AS for CSG selection	Read the HNB name from BCCH on SIB9 if a cell with a CSG ID is found
	Select a femto-CSG and request AS to select a cell belonging to this CSG	Report CSG ID of the found cell broadcasting a CSG ID together with the HeNB name and PLMNs to NAS
		Upon selection of a femto-CSG by NAS, select any cell belonging to the selected CSG fulfilling the cell selection criteria and not barred or reserved for operator use

MME/HSS. This authentication function also establishes the security keys that are used for encryption of the bearers [28]. The UE is normally paged using its S-TMSI. The *Paging Request* message also contains a UE ID index value in order for the eNB to calculate the paging occasions at which the UE will listen to paging messages (see Chapter 5 for more details).

4.4.1.3 Cell reselection

The primary purpose of cell reselection, regardless of intra-frequency, inter-frequency, or inter-RAT nature, is to ensure that the UE camps on or connects to the best cell in terms of radio condition, e.g., in terms of path loss or RSRP metric. The UE should conduct some measurements to support this procedure. A UE in RRC_IDLE state performs cell reselection. The UE measures some attributes of the serving and neighbor cells to enable the reselection process. There is no need to inform the UEs of neighboring cells in the serving cell system information to enable the UEs to search and conduct measurements on neighbor cells, i.e., E-UTRAN relies on the UE to detect the neighbor cells. For the search and measurement of inter-frequency neighbor cells, only the carrier frequencies

need to be indicated. Measurements may be avoided, if the serving cell attributes fulfill particular search or measurement criteria.

Cell reselection identifies the cell that the UE may camp on. The cell reselection procedure is based on certain criteria which involve measurements of the serving and neighbor cells. Intra-frequency cell reselection is based on ranking of the cells, whereas inter-frequency cell reselection is based on absolute priorities where a UE attempts to camp on the highest priority frequency available. Absolute priorities for reselection are provided only by the registered PLMN and are valid within the registered PLMN. These priorities are provided by the system information and are valid for all UEs in a cell. Specific priorities per UE can be signaled in the *RRC Connection Release* message. A validity time can be associated with UE-specific priorities [5].

For inter-frequency neighbor cells, it is possible to indicate layer-specific cell reselection parameters (e.g., layer-specific offset). These parameters are common to all neighbor cells on a frequency. A neighbor cell list can be provided by the serving cell to assist specific cases for intra- and inter-frequency neighbor cells. This neighbor cell list contains cell-specific cell reselection parameters (e.g., cell-specific offset) for specific neighbor cells. Black lists can also be provided to prevent the UE from reselecting specific intra- and inter-frequency neighbor cells. Cell reselection can be speed-dependent (speed detection based on UTRAN solution). Cell reselection parameters are applicable for all UEs in a cell, but it is possible to configure specific reselection parameters per UE group or per UE. Cell access restrictions apply similar to UTRAN, which consist of access class barring and cell reservation (e.g., for cells "reserved for operator use") which are applicable for the UEs in RRC_IDLE mode [5,20].

4.4.1.4 Idle mode DRX

E-UTRA supports DRX in the downlink in order to reduce UE power consumption during periods of inactivity. In the DRX mode, the UE monitors only PDCCH during a predefined on-duration interval within one of two possible DRX cycles. The UE may turn off its receiver when it does not expect to receive any transmission in the downlink over PDCCH. The UE can transition to DRX mode in one of the following ways: (1) implicitly based on the expiration of certain configured timers; or (2) explicitly based on the transmission of a DRX command in the form of a MAC control element. A UE can enter the DRX mode of operation during the active mode (DRX in RRC_CONNECTED state) as well as in the idle mode (DRX in RRC_IDLE state). The DRX in the connected state saves UE battery power, especially when the traffic is characterized by long periods of inactivity such as web-browsing traffic. This is achieved by interleaving short periods of active monitoring of the PDCCH with longer periods of inactivity. While in the idle mode utilizing DRX, the UE may wake up periodically and check the paging channel to determine whether it has any pending downlink traffic. Generally, an active mode DRX interval is shorter than an idle mode DRX interval.

It is important to note that E-UTRA supports DRX operation in both idle and connected modes. The DRX in RRC_IDLE state is used when monitoring the paging messages during paging occasions, whereas the DRX in RRC_CONNECTED state is used to conserve power and determines the subframes during which the UE monitors PDCCH using C-RNTI, TPC-PUCCH-RNTI, TPC-PUSCH-RNTI, SI-RNTI, or SPS C-RNTI for a possible data/control message. When idle mode DRX is used, the UE needs to monitor only one P-RNTI per DRX cycle.

There are a number of timers associated with the DRX mode. The DRX inactivity timer controls the number of subframes for which a UE continues to monitor the downlink before the absence of activity triggers the UE to enter the DRX mode. Upon entering the DRX mode, the UE monitors the downlink for a number of subframes defined by the on-duration timer in every short DRX cycle or long DRX cycle. The short DRX cycle is optional. If it is configured, the UE will follow the short DRX cycle for a specific number of cycles controlled by the short DRX cycle timer before switching to the long DRX cycle, if there is no activity. Other timers prevent the UE from entering the DRX mode. These are associated with HARQ downlink retransmissions. If the UE has received data that it was unable to decode and this was not associated with a broadcast HARQ process, then following the expiration of the HARQ round-trip-time timer, which specifies the earliest time at which a UE can expect a retransmission, the DRX retransmission timer is started. The UE must wait for this timer to expire and continue to monitor the downlink before it can enter the DRX mode. In addition to the above timers, the network can order the UE to enter the DRX mode with the DRX command MAC control element. All timers and parameters described in this context are configured by RRC signaling.

During the DRX procedure, the UE maintains a DRX cycle that is defined as a number of subframes. The UE monitors PDCCH for a specific number of subframes in a DRX cycle. This time interval is called on-duration and can range from 1 to 200 subframes. The UE may turn off its receiver for the rest of the DRX cycle. The UE maintains two DRX cycles referred to as short DRX cycle and long DRX cycle, which have different durations. The short DRX cycle is optional and targeted for applications such as VoIP which typically require relatively small transmissions of data at short but regular intervals. If configured, the UE starts with a short DRX cycle (2−640 subframes) when it enters the DRX mode. When the configurable short-DRX timer expires, the UE transitions into the long DRX cycle (10−2560 subframes). The UE can transition to the DRX mode either upon expiration of configured timers or following a DRX command in the form of a MAC control element.

4.4.2 Connected mode procedures

The RRC_CONNECTED is a state where the UE transfers unicast control/data in the uplink or downlink. The UE may be configured with an individualized DRX or discontinuous transmission pattern, or both. The UE monitors control channels

associated with the shared data channel to determine if data is scheduled for it, provides channel quality and feedback information, performs neighbor cell measurements and measurement reporting, and acquires the system information. One or multiple IP addresses are assigned to the UE as well as a temporary identifier called C-RNTI that is used for signaling purposes between the UE and the network. In brief, the RRC_CONNECTED state is characterized as follows:

- UE is known in EPC and E-UTRAN/eNB, i.e., UE context is stored in eNB.
- UE location is known at cell level.
- Unicast data transfer with UE is possible.
- Short/long DRX cycles are supported for power saving.
- Mobility is controlled by the network.

In RRC_CONNECTED state, the network controls UE mobility, i.e., the network decides when the UE must move to another cell, which may be on another frequency or RAT. For network-controlled mobility in RRC_CONNECTED state, handover is the only procedure that is defined. The network triggers the handover procedure based on radio conditions or network load. The network may configure the UE to perform measurement reporting, including the configuration of measurement gaps. The network may also initiate handover without receiving measurement reports from the UE. Before sending the handover message to the UE, the source eNB prepares one or more target cells. The target eNB generates the message used to perform the handover including the AS configuration to be used in the target cell. The source eNB transparently forwards the handover message/information received from the target to the UE. The source eNB may initiate data forwarding for some of the user data radio bearers (DRBs). After receiving the handover message, the UE attempts to access the target cell at the first available random access channel opportunity according to the random access resource selection procedure for asynchronous handover. When allocating a dedicated preamble for the random access in the target cell, the E-UTRA ensures it is available from the first random access opportunity the UE may utilize. Upon successful completion of the handover, the UE sends a message used to confirm the handover. Following successful completion of handover, the PDCP SDUs may be retransmitted in the target cell. This only applies to user DRBs using RLC acknowledged mode. Furthermore, the sequence number and the hyper-frame number (HFN)[4] are reset except for the user DRBs using RLC acknowledged mode, for which both sequence number and HFN continue. For a limited period of time, the source eNB maintains the UE context to allow the UE to return in case of handover failure. If handover failure is detected, the UE will attempt to resume the RRC connection either with the source eNB or with another cell using

[4]An HFN is an overflow counter mechanism that is used in the eNB and UE in order to limit the actual number of sequence number bits that are needed to be sent over the radio air-interface. The HFN needs to be synchronized between the UE and eNB.

the RRC connection reestablishment procedure. The connection reestablishment would succeed, if the accessed cell is prepared [3].

The paging procedure is used to transmit paging information to a UE in RRC_IDLE and/or to inform mobile terminals in RRC_IDLE or RRC_CONNECTED states about an system information change and/or to inform about an ETWS primary and/or secondary notification. The paging information is provided to upper layers, which in response may establish an RRC connection to receive an incoming call. The E-UTRAN initiates the paging procedure by transmitting the paging message at the UE's paging opportunities. The E-UTRAN may address multiple mobile terminals within a paging message by including one paging record for each UE. The E-UTRAN may also indicate a change of system information and/or provide an ETWS notification.

The RRC connection establishment involves SRB1 bearer establishment. The procedure is also used to transfer the initial NAS dedicated information/message from the UE to the E-UTRAN. The UE initiates the RRC connection establishment procedure when the upper layers request establishment of an RRC connection while the UE is in the RRC_IDLE state.

The RRC connection reestablishment is used to reestablish the RRC connection including resumption of SRB1 operation and the reactivation of security. A UE in RRC_CONNECTED state, for which security has been activated, may initiate the procedure in order to continue the RRC connection. The connection reestablishment succeeds if the serving/target node is prepared, i.e., it has a valid UE context. If the E-UTRAN accepts the connection reestablishment, SRB1 operation resumes while the operation of other radio bearers are suspended. If the AS security has not been established, the UE cannot initiate the procedure and directly transitions to the RRC_IDLE state [3].

4.5 Network attach procedures

After establishing an RRC connection, the UE can send an *Attach Request* message to the MME in order to register with the network. The UE is required to further request PDN connectivity along with the network attach request. After all necessary signaling connections are established, the EPC may trigger security functions. When a UE connects to the LTE radio access network, it is automatically connected to the PDN and this connected state is continuously maintained. In other words, as the terminal is registered in the network or alternatively attached to the core network through the LTE radio access network, a communication path to the PDN (providing IP connectivity) is established. The PDN to which a connection is established can be configured on a per-subscriber basis, or the terminal can select it during the attach procedure. This PDN is called the default PDN. With the always-on connection function, only the radio link part of the connection is released after a certain time interval elapses with no terminal activity; however,

the IP connectivity between the terminal and the network is sustained. By doing so, only the radio link needs to be reconfigured when the terminal resumes data communication in order to reduce the connection setup time. Furthermore, the IP address obtained when the terminal attaches can be used until it detaches, so it is always possible to receive packets using that IP address [29,35].

The procedure for a UE when initially attaching to the core network until connection to the PDN is established is shown in Figure 4.10. The details of attach procedure can be described as follows [25]:

1. The UE establishes an RRC connection with the eNB.
2. The UE sends the *Attach Request* message together with a *PDN Connectivity Request* message for PDN (IP) connectivity through the established RRC connection. As part of this process, the eNB establishes an S1 logical connection with the MME for the UE. In other words, when the UE establishes an RRC connection with the eNB for sending and receiving RRC and NAS messages, it sends an *Attach Request* message to the MME. The UE and the MME perform the required security procedures, including authentication, encryption, and integrity protection. The purpose of the attach request is to register with the EPS network, whereas the purpose of the PDN connectivity request is to establish connectivity with the PDN for transporting user data. In order to transport NAS messages, the UE requests AS to establish NAS signaling connection.
3. If the UE has valid security parameters, the *Attach Request* message must be integrity protected in order to allow validation of the UE by the MME, KSI_{ASME}, NAS sequence number, and NAS message authentication code are included if the UE has valid EPS security parameters. NAS sequence number indicates the sequential number of the NAS message. If the UE does not have a valid EPS security association, then the *Attach Request* message is not integrity protected. The UE network capabilities also indicate the supported NAS and AS security algorithms. The PDN type indicates the requested IP version (IPv4 or IPv6 [17]). If the network is not able to identify the UE with the ID given in the *Attach Request* message, it initiates the identification followed by authentication and security activation procedures.
4. A single eNB can be associated with multiple MMEs within a single PLMN or across multiple PLMNs. Furthermore, the eNB does not maintain any information about the UE in the RRC_IDLE state. During initial attach, the UE is not registered with any MME. As a result, the eNB selects an MME in the selected PLMN based on the load status of MMEs. The load information is provided by the MME(s) on a periodic basis. If an eNB only connects to a single MME, then the MME selection will not be required. An MME pool area is defined as an area consisting of a number of tracking areas. This is an area within which a UE may be served without changing the serving MME. In this pool of MMEs, the eNB can be connected to multiple MMEs and has a choice to select any MMEs. The load balancing among MMEs enables the

Samsung Ex. 1005, Page 212 of 1147

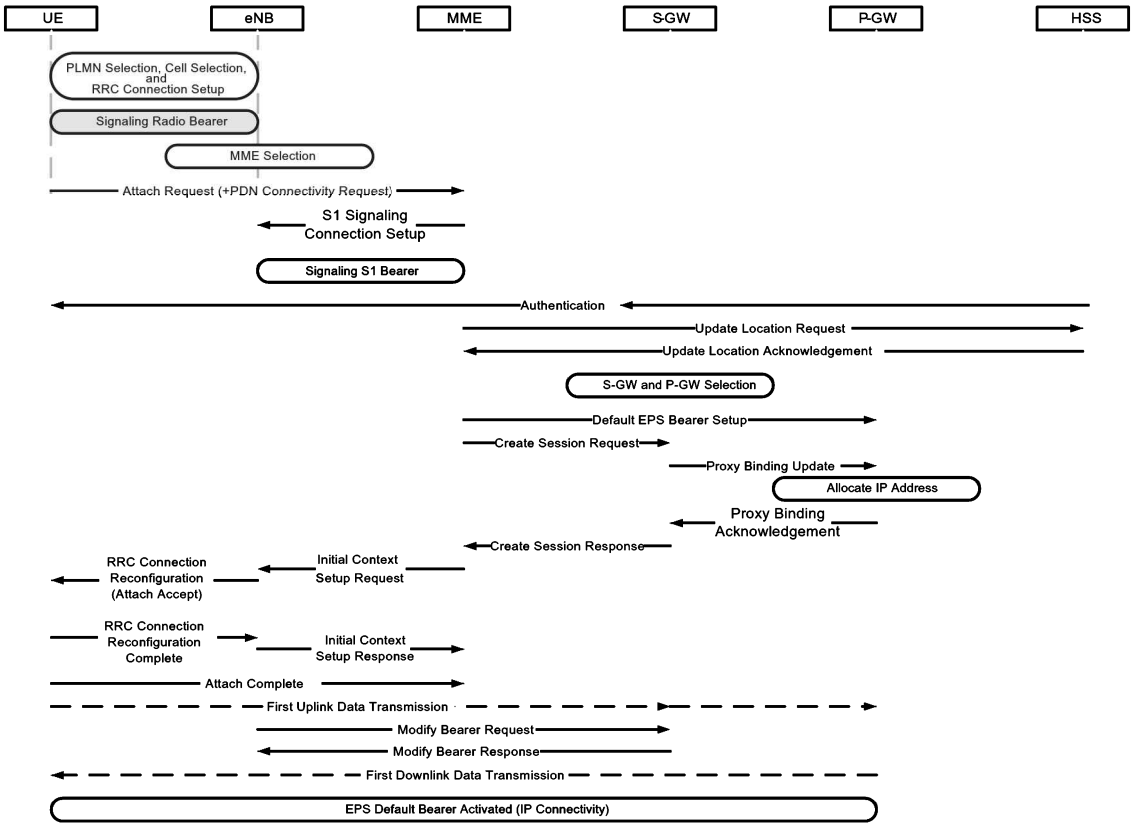


FIGURE 4.10

Attach procedure [10,28,29].

network to ensure equally loaded MMEs within a pool area. This is achieved by setting a weight factor for each MME such that the probability of selecting an MME is proportional to its weight factor. The weight factor is typically set according to the capacity of an MME relative to other MMEs.

5. The eNB and MME are connected via an S1-MME interface and use S1 application protocol [12]. An S1-MME interface is established when the eNB and MME connect with each other during initial provisioning.
6. The MME updates the home subscriber server with the location of the UE using the *Update Location Request* message based on the Diameter protocol.[5] It also requests the subscriber profile from the HSS using this message. The HSS updates its database with the current location of the UE and sends the subscriber profile information to the MME in the Diameter *Update Location Acknowledgement* message.
7. There can be multiple P-GWs and S-GWs for a given access point name (APN) within a single PLMN; only one S-GW and one P-GW serve a UE with an APN. After subscriber data is downloaded, the MME initiates EPS-bearer setup with P-GW. To accomplish this, the MME selects a P-GW in the subscribed default APN and then selects an S-GW. Note that a wireless network connects with multiple external networks, including the Internet and specific enterprise networks. An APN identifies these external networks during a data session. The APN specifies the external networks that a mobile device can access. It also defines which type of IP address to use, which security mechanisms to invoke, and which fixed-end connections to use (fixed-end connections refer to the connections between an operator's wireless network and the customer network). The APNs available to a device are specified as part of the subscriber account. A mobile station can have access to more than one APN, where each packet data protocol context (data session) specifies which individual APN to use.
8. To establish a transmission path to the default PDN, the MME sends a *Create Session Request* message to the S-GW.
9. When the S-GW receives the create session request from the MME, it requests proxy binding update from the P-GW, i.e., a request message sent by a mobile access gateway to a mobile node's local mobility anchor for establishing a binding between the mobile node's home network prefixes assigned to a given interface of a mobile node and its current care-of address. The P-GW allocates an IP address to the terminal and notifies the

[5]The Diameter protocol is designed to provide an authentication, authorization, and accounting (AAA) framework for applications such as network access or IP mobility. Diameter is also intended to work in local AAA as well as roaming scenarios. Diameter protocol was designed as an improved version of the RADIUS protocol; however, it does not address flaws within the RADIUS model. The basic concept behind Diameter is to provide a base protocol that can be extended in order to provide AAA services to new access technologies. Diameter does not use the same RADIUS protocol data unit but is backward compatible with RADIUS to facilitate network migration [19].

Samsung Ex. 1005, Page 214 of 1147

S-GW of this information in a *Proxy Binding Acknowledgement* message, i.e., a reply message sent by a local mobility anchor in response to a *Proxy Binding Update* message that it received from a mobile access gateway. This process establishes a continuous core network communication path between the P-GW and the S-GW for the allocated IP address [30].

10. A radio access bearer is established from S-GW to the eNB, where a create session response message is sent to the MME. The *Create Session Response* message contains information required to configure the radio access bearer from the eNB to the S-GW, including information elements issued by the S-GW and the IP address allocated to the terminal.
11. The MME sends the information in the *Create Session Response* message to the eNB in an initial context setup request message. Note that this procedure also contains other notifications such as *Attach Accept*, which is the response to the *Attach Request*. When the terminal receives *Attach Accept*, it sends an *Attach Complete* response to the MME, notifying that the process is complete.
12. The eNB establishes the radio data link and sends the *Attach Accept* message to the terminal. It also configures the radio access bearer from the eNB to the S-GW and sends an initial context setup response to the MME. The initial context setup response contains information elements issued by the eNB required to establish the radio access bearer from the S-GW to the eNB.
13. The MME sends the information in the initial context setup response to the S-GW in a modify bearer request message. The S-GW completes configuration of the previously prepared radio bearer from the S-GW to the eNB and sends a modify bearer response to the MME.

Through the above steps, a communication path from the terminal to the P-GW is established, enabling communication with the default PDN. If the terminal has no activity for a certain period of time, the always-on connection function described above releases the radio control link and the LTE radio access bearer, while maintaining the core network communication path.

After the terminal has established a connection to the default PDN, it is possible to initiate another connection to a different PDN. In this way, it is possible to manage PDNs according to service agreements. For example, the IMS PDN, which provides voice services through a packet network, may be used as the default PDN, and a different PDN may be used for Internet access. To establish a connection to a PDN other than the default PDN, the procedure is the same as the attach procedure shown in Figure 4.10, except no location update is required.

Once the attach procedure succeeds, a context is established for the UE in the MME, a default bearer is set up between the UE and the P-GW, and an IP address is assigned to the UE. At this time, the UE has IP connectivity and it can start using IP-based Internet or IMS services.

If the UE does not require services from the network, e.g., when the UE is switched off, it needs to deregister itself from the network by initiating the detach procedure. The detach procedure is initiated by the UE by sending a *Detach*

Request message. The detach type information element included in the message indicates whether detach is due to a switch off. The network and the UE deactivate the EPS-bearer context(s) for this UE locally without additional signaling between the UE and MME. If the detach type is "switch off," then the MME does not send the *Detach Accept* message; otherwise, the message is sent to the UE. While processing the detach message from the UE, the MME initiates the release of the EPS bearers in the network and it also flushes the UE context held at the eNB.

4.5.1 Random access procedure

A UE may access network services once it is synchronized in downlink and uplink directions. After the PLMN and cell selection phase, the UE is synchronized with the network in the downlink direction. Subsequently, it needs to synchronize with the network in the uplink direction. The random access procedure over PRACH is performed by the UE for this purpose. The random access procedure is characterized as one procedure independent of cell size and is the same for FDD and TDD systems. There are two types of random access procedures:

- *Contention-based random access procedure*: In this case, multiple UEs may simultaneously attempt to access the network, resulting in potential collisions. The contention is resolved among the contending UEs using the following procedure [3]:
  1. Transmission of random access preamble on RACH in the uplink. The RACH preambles are grouped based on the size of the L3 message that the UE would like to transmit; this provides an indication of resource requirements for the UE to the network.
  2. Transmission of random access response generated by eNB MAC on DL-SCH. This is an indication to the UE that the eNB has received its preamble and conveys the resources reserved for this UE.
  3. Transmission of the first scheduled uplink transmission on UL-SCH. The UE sends the *RRC Connection Request* message using the resources allocated by the eNB in the previous step. In the *RRC Connection Request* message, the UE sends its identifier to the network which is used by the eNB to resolve the contention in the next step.
  4. Contention resolution in the downlink. The eNB sends back the received UE identifier to resolve the contention, and at this point the UE which has received its identifier continues with the transmission while others will back off and try again.
- *Non-contention-based random access procedure*: The network initiates this procedure in the case of a handover of a UE from one eNB to another in order to reduce the handover latency. The network usually reserves a set of RACH preambles for this purpose, where one is assigned and sent to the UE. There are no collisions with other UEs' preamble transmissions in the uplink

because the eNB controls the process. Once the UE receives the assigned random access preamble, it sends it to the eNB. The eNB then generates and sends a *Random Access Response* message.

As shown in Figure 4.11, during the first step of the contention-based procedure, the UE transmits a preamble randomly selected from 64 possible preamble sequences which are divided into two groups. The group from which the UE selects the preamble depends on the size of the message it would like to transmit and on the calculated path loss to the eNB. Once the preamble is transmitted, the UE monitors the downlink beginning three subframes after the end of the preamble transmission for a duration controlled by a configurable window. If no downlink response is detected, the preamble is retransmitted with power increased by a configurable parameter. This process continues until the maximum number of random access preamble transmissions is reached. If the preamble is successfully detected by the eNB, then a *Random Access Response* message is sent in the downlink, allocating resources for the UE to transmit its message or enforcing a backoff. In the final step, contention resolution must be performed for a contention-based procedure.

The UE sets its initial power level to the following value *preambleInitialReceivedTargetPower* − *pathloss* + *DELTA_PREAMBLE*, where *DELTA_PREAMBLE* is a fixed value that depends on the type of preamble format. The UE selects the first available RACH resource for the transmission. If there is no response within three subframes + *ra-ResponseWindowSize* (defined in number of subframes), the UE assumes that the transmission was unsuccessful. It selects another RACH preamble, performs the backoff procedure, and transmits the new preamble. The power level is increased by *powerRampingStep*. This process can be repeated until *preambleTransMax* preambles are transmitted, in which case a failure is declared to the upper layers.

Once the random access preamble is detected by the eNB, it responds with a *Random Access Response* message. This message provides information that the UE needs to transmit in the uplink including the timing advance and the grant specifying which resources to utilize. To resolve a possible contention, the UE sends an *RRC Connection Request* message using a random number (or the S-TMSI if available) as its identity. A contention resolution timer is started and is monitored in the downlink for the *Contention Resolution* message. If this message is received and contains the UE ID transmitted in step 3, the contention is resolved. The temporary C-RNTI contained in the *Random Access Response* message becomes the C-RNTI. If the *Contention Resolution* message containing the correct UE identity is not received before the contention resolution timer[6] expires, the UE assumes the access failed and the random access procedure needs to be repeated.

[6]The contention resolution timer specifies the number of consecutive subframe(s) during which the UE monitors the physical downlink control channel after Msg3 is transmitted [6].

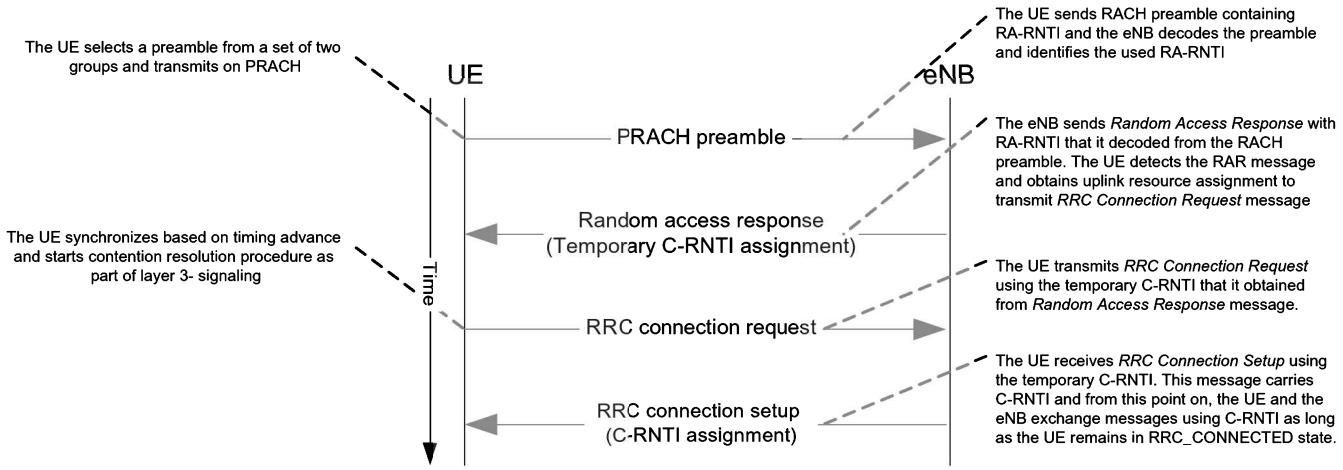


FIGURE 4.11

Contention-based random access procedure [3,25,32].

4.5.2 RRC connection establishment, modification, and release

One of the main functions of the RRC sublayer is to establish, maintain, modify, and release connections with E-UTRAN for transport of user data and control messages. The RRC connection establishment starts with the establishment of signaling radio bearer 1 (SRB1) and the transfer of the initial NAS message. This NAS message triggers the setup of an S1 connection, which normally initiates a subsequent step during which E-UTRAN activates AS security and further establishes signaling radio bearer 2 (SRB2) along with one or more DRBs associated with the default and optionally dedicated EPS bearers. Figure 4.12 illustrates the RRC connection establishment, NAS signaling, and RRC control messages. As shown in the figure, the RRC connection establishment procedure consists of three steps: the *RRC Connection Request* message from the UE; the *RRC Connection Setup* message from the eNB; and the *RRC Connection Setup Complete* message from the UE. Upon successful completion of the RRC connection establishment procedure, the UE is in the RRC_CONNECTED state and is provided with periodic timing alignment commands by the eNB. The UE provides channel quality feedback to the eNB and conducts neighboring cell measurements for handover. Note that NAS messages (ECM/EMM) are encapsulated in RRC messages; thus RRC connection establishment needs to happen prior to NAS signaling transfer. Furthermore, it must be noted that signaling radio bearers (SRBs) are established before DRBs. Upon receiving the UE context from the EPC, the E-UTRAN activates security that includes ciphering and integrity protection using the initial security activation procedure.

It must be noted that an SRB is defined as a radio bearer that is only used for transmission of RRC and NAS messages. More specifically, the following three SRB types are defined in E-UTRA [3,9]:

1. SRB0 is used for transport of RRC messages associated with the common control logical channel.
2. SRB1 is used for transport of RRC messages including piggybacked NAS messages as well as NAS messages prior to the establishment of SRB2 where all are associated with the dedicated control logical channel.
3. SRB2 is used for transport of NAS messages using the dedicated control logical channel, has a lower priority than SRB1, and is always configured by E-UTRAN after security activation.

The UE in the RRC_IDLE state does not have any allocated SRB or DRB. The UE requires (dedicated) SRBs and DRBs to exchange data/messages with the network. In particular, for transfer of NAS messages, SRBs are established. They carry registration and signaling for user-plane setup. The user-plane application data is transported on DRBs. An E-UTRAN radio access bearer (E-RAB)

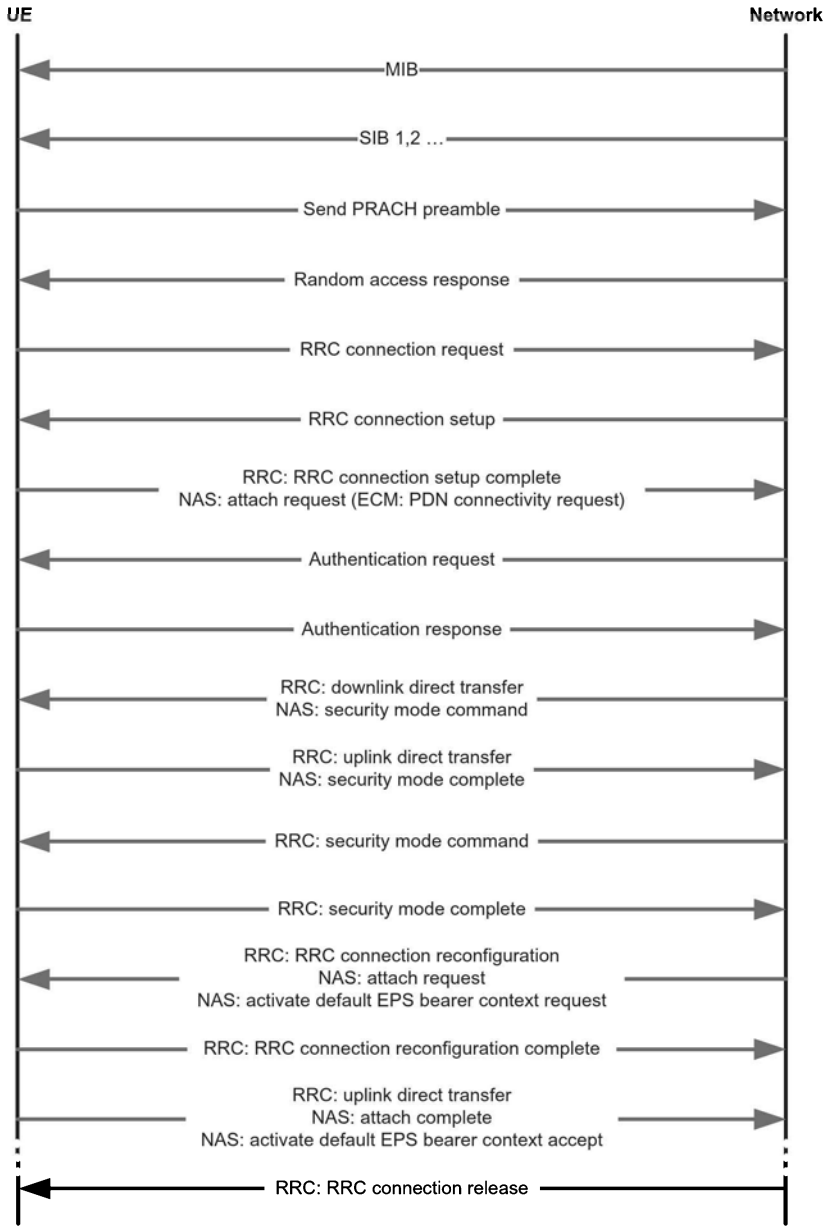


FIGURE 4.12

Sequence of signaling for RRC connection setup, reconfiguration, and release [32].

uniquely identifies the concatenation of an S1 bearer and the corresponding DRB. When an E-RAB exists, there is a one-to-one mapping between the E-RAB and an EPS bearer. To establish, modify, or release DRBs, the E-UTRAN applies the RRC connection reconfiguration procedure.

To allow a wide range of UE implementations, different UE capabilities are specified in the E-UTRA. The UE capabilities are divided into a number of parameters, which are sent from the UE at the establishment of a connection and if/when the UE capabilities are changed during an ongoing connection. The UE capabilities may then be used by the network to select a configuration that is supported by the UE.

In E-UTRAN, the UE radio network-related capability information are transferred using RRC signaling from the UE to the eNB. In order to avoid sending the UE capabilities over the radio interface between the UE and the eNB each time that the UE performs a transition to the connected mode (i.e., when the UE-specific context is created in the eNB), the eNB stores the UE capabilities in the MME so that they are safeguarded when the UE is in the idle mode. When the UE returns to the connected mode, the UE capabilities should be downloaded to the eNB. Before the UE can fully use the services of the network, it has to declare its capabilities (e.g., supported bit rates, antenna configurations, bandwidths, and supported radio access types) to the network. When the network has knowledge of the UE capabilities, it can optimize transmissions in conjunction with services to the UE according to its capabilities.

In general, the UE capabilities can be classified into mainly two categories depending on which layer of the protocol hierarchy the given capability is related to. The access stratum-related capabilities are the access technology-dependent parts of the capability information, such as terminal power class and supported frequency band. The AS capabilities are required by the eNB. The non-access stratum-related capabilities are a set of capability information containing the non-access parts of the UE capability such as supported security algorithms. The NAS capabilities are used by the EPC.

The AS part of the UE capabilities needs to be present in the eNB, while the UE is in the connected mode. Moreover, when a handover is made from the source eNB to the target eNB, the UE capability information needs to be transferred from the source eNB to the target eNB. However, for a UE in the idle mode, there is no need to maintain any UE information, including the UE capabilities, in the eNBs, but only in the MME. Therefore, when a UE transitions to the active state again, the information of the UE, including the UE capabilities, must be recreated in the eNB [20,21].

The *UE Capability Enquiry* message is used to request the transfer of UE radio access capabilities for E-UTRA as well as for other RATs (UTRA, GERAN-CS, GERAN-PS, and CDMA2000). As the UE's AS capability may contain several RATs capabilities, the information can be very large. To reduce the overhead on the air-interface during transition from RRC_IDLE state to RRC_CONNECTED state, the MME stores the UE's AS capability and provides

it to the eNB during initial UE context setup over the S1 interface. If the MME does not have a valid UE's AS capability, it does not provide UE's AS capability to the eNB. The eNB has to acquire the UE's AS capability from the UE using a *UE Capability Enquiry* message. The eNB sends a *UE Capability Enquiry* message to transfer the UE's radio access capability information to the E-UTRAN. The eNB identifies the RAT for which it is requesting the capabilities. This message supports requests for E-UTRAN, UTRAN, GERAN, and CDMA2000 1xRTT. The UE responds with a *UE Capability Information* message, which includes the requested RAT container(s), provided that the requested RAT by the eNB is supported. The information obtained is used to configure MAC and physical-layer attributes (receive and transmit capabilities such as single/dual radio, and dual receiver) of the connection. It also enables efficient measurement control, preventing unnecessary UE wake-ups to conduct measurements [9].

4.5.3 Service establishment and QoS

There are two types of EPS bearers in E-UTRA. One is called "default EPS bearer" and the other is known as "dedicated EPS bearer," which can be described as follows. The default EPS bearer is established during the initial attach procedure to allocate an IP address to the UE, and does not have a specific QoS (only nominal QoS is applied). This is similar to the primary PDP context in UMTS [24,32,36]. The dedicated EPS bearer is typically established during a call setup following transition from the idle mode, which does not allocate any additional IP address to the UE and is linked to a specified default EPS bearer. However, it has a specific and usually guaranteed QoS. This is similar to the secondary PDP context in UMTS [24,28].

An EPS bearer is the level of granularity for bearer-level QoS control in the EPC/E-UTRAN. It means that the service data flows (SDFs), which are mapped to the same EPS bearer, receive the same bearer-level packet forwarding treatment (e.g., scheduling policy, queue management policy, rate shaping policy, and RLC configuration). One EPS bearer is established when the UE connects to a PDN and is sustained throughout the lifetime of the PDN connection to provide the UE with always-on IP connectivity to that PDN. This bearer is referred to as the *default* bearer. Any additional EPS bearer that is established to the same PDN is referred to as a *dedicated* bearer. The initial bearer-level QoS parameter values of the default bearer are assigned by the network, based on subscription data. The decision to establish or modify a dedicated bearer can only be taken by the EPC, and the bearer-level QoS parameter values are always assigned by the EPC [20].

An EPS bearer is referred to as a guaranteed bit rate (GBR) bearer, if dedicated network resources related to a GBR value that is associated with the EPS bearer are permanently allocated (e.g., by an admission control function in the eNB) during bearer establishment or modification. Otherwise, an EPS bearer is

referred to as a non-GBR bearer. A dedicated bearer can either be a GBR or a non-GBR bearer, while a default bearer is a non-GBR bearer.

The GBR bearers are mainly used to carry voice, video, and real-time gaming services through dedicated bearers or static scheduling. Each GBR bearer is associated with the bit rate that is expected to be provided by the bearer, while the maximum bit rate (MBR) indicates the upper limit of the bit rate that can be provided by that bearer. Non-GBR bearers are mainly used to carry various data services. To improve bandwidth utilization, EPS introduces the concept of aggregated MBR (AMBR). AMBR is an IP connectivity access network (IP-CAN), which refers to an access network that provides IP connectivity, session level QoS parameter of every PDN connection. Multiple EPS bearers for the same PDN connection share the same AMBR value. Each non-GBR bearer can potentially make use of the whole AMBR if other EPS bearers do not transfer any services. Therefore, an AMBR actually restricts the total bit rate of all the bearers sharing this AMBR [24].

The EPS-bearer service layout is depicted in Figure 4.13. An uplink traffic flow template (TFT) in the UE binds an SDF to an EPS bearer in the uplink direction. Multiple SDFs can be multiplexed into the same EPS bearer by including multiple uplink packet filters in the uplink TFT. A downlink TFT in the P-GW binds an SDF to an EPS bearer in the downlink direction. Similar to the uplink, multiple SDFs can be multiplexed into the same EPS bearer by including multiple downlink packet filters in the downlink TFT. An E-RAB transports the packets of an EPS bearer between the UE and the EPC. When an E-RAB is established, there is a one-to-one mapping between the E-RAB and an EPS bearer. A DRB transports the packets of an EPS bearer between a UE and an eNB. When a DRB exists, there is a one-to-one mapping between this DRB and the EPS bearer/E-RAB [10].

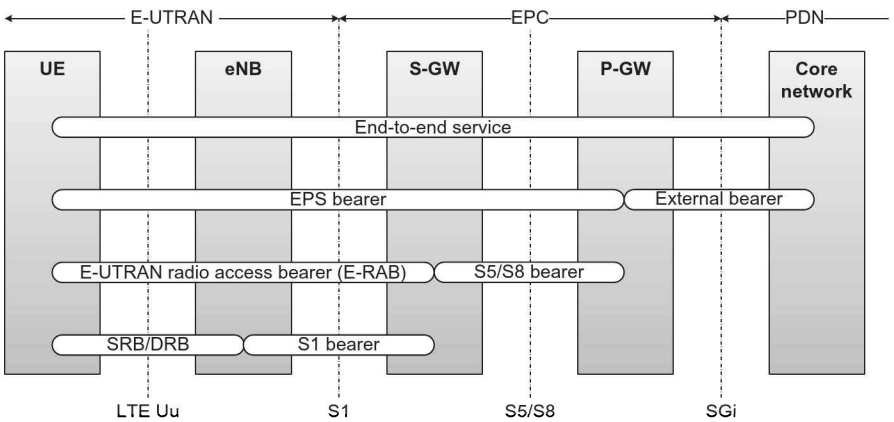


FIGURE 4.13

EPS-bearer service structure [3].

The bearer level (i.e., per bearer or per aggregate) QoS parameters are QCI, ARP, GBR, and AMBR [24]. Each EPS bearer/E-RAB (GBR and non-GBR) is associated with the following bearer level QoS parameters [3]:

- QoS class identifier (QCI): A scalar that is used as a reference to access node-specific parameters that control bearer-level packet forwarding behavior (e.g., scheduling weights, admission thresholds, queue management thresholds, and link layer protocol configuration), which have been preconfigured by the operator. A one-to-one mapping of QCI values to standardized characteristics is captured in 3GPP TS 23.401 [10].
- Allocation and retention priority (ARP): The primary purpose of ARP is to decide whether a bearer establishment/modification request can be accepted or needs to be rejected in case of resource limitation. In addition, the ARP can be used by the eNB to decide which bearer(s) to drop during resource limitation conditions such as handover. The ARP priority level ranges from 1 to 15, where 1 is the highest level of priority.

Each GBR bearer is additionally associated with the following bearer level QoS parameters [3]:

- Guaranteed bit rate (GBR): The bit rate that is expected to be provided by a GBR bearer.
- Maximum bit rate (MBR): The maximum bit rate that is expected to be provided by a GBR bearer. The MBR can be greater than or equal to the GBR.

Each APN access by a UE is associated with per APN aggregate maximum bit rate (APN-AMBR) QoS parameter. The APN-AMBR limits the aggregate bit rate across all non-GBR bearers and across all PDN connections of the same APN (excess traffic may be discarded). Each UE in the EMM-REGISTERED state is associated with a per UE aggregate maximum bit rate (UE-AMBR) QoS parameter. The GBR and MBR denote the bit rate of traffic per bearer, while UE-AMBR/APN-AMBR denote the bit rate of traffic per group of bearers. The UE-AMBR limits the aggregate bit rate across all non-GBR bearers of a UE (excess traffic may be discarded by a rate shaping function). Each of these QoS parameters has an uplink and a downlink component.

UE-AMBR and ARP are network parameters that are not known at the UE. The standardized QCI characteristics are not signaled on any interface. They are guidelines for the preconfiguration of node-specific parameters for each QCI. Standardizing a QCI with corresponding characteristics ensures that applications/services mapped to that QCI receive the same minimum level of QoS in multi-vendor network deployments and in case of roaming. Each QCI (GBR or non-GBR) is associated with a priority level, with priority level one being the highest priority level. Scheduling between different SDF aggregates are primarily based on the packet delay budget (PDB), which refers to the permissible latency of data packets transported between UE and P-GW. For E-UTRAN, the priority level of a QCI may be used as the basis for assigning the uplink priority per radio bearer. The purpose of

the PDB is to support the configuration of scheduling and link layer functions (e.g., setting scheduling priority weights, and HARQ target operating points). For a specific QCI, the PDB is the same for both uplink and downlink. Packet loss rate (PLR) is defined as the rate of SDUs not successfully transported to the upper layer to the total number of SDUs processed by the link layer of the transmitting end. Therefore, PLR reflects the upper bound of tolerable packet loss rate in non-congested cases and is identical in both uplink and downlink for the same QCI [36].

4.6 Data call establishment procedures

4.6.1 UE-originated data call procedure

After successfully attaching to the network, the UE can request services from the network using a service request procedure. When the UE requests resources from the network to initiate a data call, it can utilize a NAS service request procedure. Another example of this procedure is during initiation of mobile-originated or mobile-terminated fallback procedures, if they are supported by the network. The call flow for this procedure is shown in Figure 4.14. This procedure can be described as follows [25]:

1. The UE establishes an RRC connection with the eNB.
2. The UE sends a service request message to the MME to request (dedicated) bearer resources by including the bearer resource allocation request. The eNB establishes an S1 logical connection with the MME for this UE. Note that the UE may also send the bearer resource allocation request to the MME as a stand-alone message. At this point, the network can initiate an optional identification followed by authentication and security mode procedures.
3. After the completion of the authentication and security mode procedure, the MME initiates activation of the default bearer with the S-GW/P-GW by sending the eGTP-C[7] modify bearer request message to the S-GW. Upon

[7]GPRS tunneling protocol (GTP) is a group of IP-based communications protocols used to carry general packet radio service (GPRS) within GSM, UMTS, and LTE networks [2]. In 3GPP architecture, GTP and Proxy Mobile IPv6-based interfaces are specified on various reference points. GTP can be decomposed into separate protocols: GTP-C; GTP-U; and GTP'. GTP-C is used within the GPRS core network for signaling between gateway GPRS support nodes (GGSN) and serving GPRS support nodes (SGSN). This allows the SGSN to activate a session on a user's behalf (PDP context activation), to deactivate the same session, to adjust QoS parameters, or to update a session for a subscriber joined from another SGSN. GTP-U is used for carrying user data within the GPRS core network and between the radio access network and the core network. The user data transported can be packets in any IPv4, IPv6, or PPP formats. The primary bearer-plane in LTE across the access and the core networks is the evolved GPRS tunneling protocol (eGTP). The user-plane variant of this protocol, eGTP-U, provides tunneling support for user data between the eNB and the EPC elements. The eGTP-C (GTPv2-C) protocol is responsible for creating, maintaining, and deleting tunnels on multiple Sx interfaces. It is used for control-plane path management, tunnel management, and mobility management. It also controls forwarding relocation messages, serving radio network subsystem (SRNS) context, and creating forward tunnels during inter-LTE handovers [2,27].

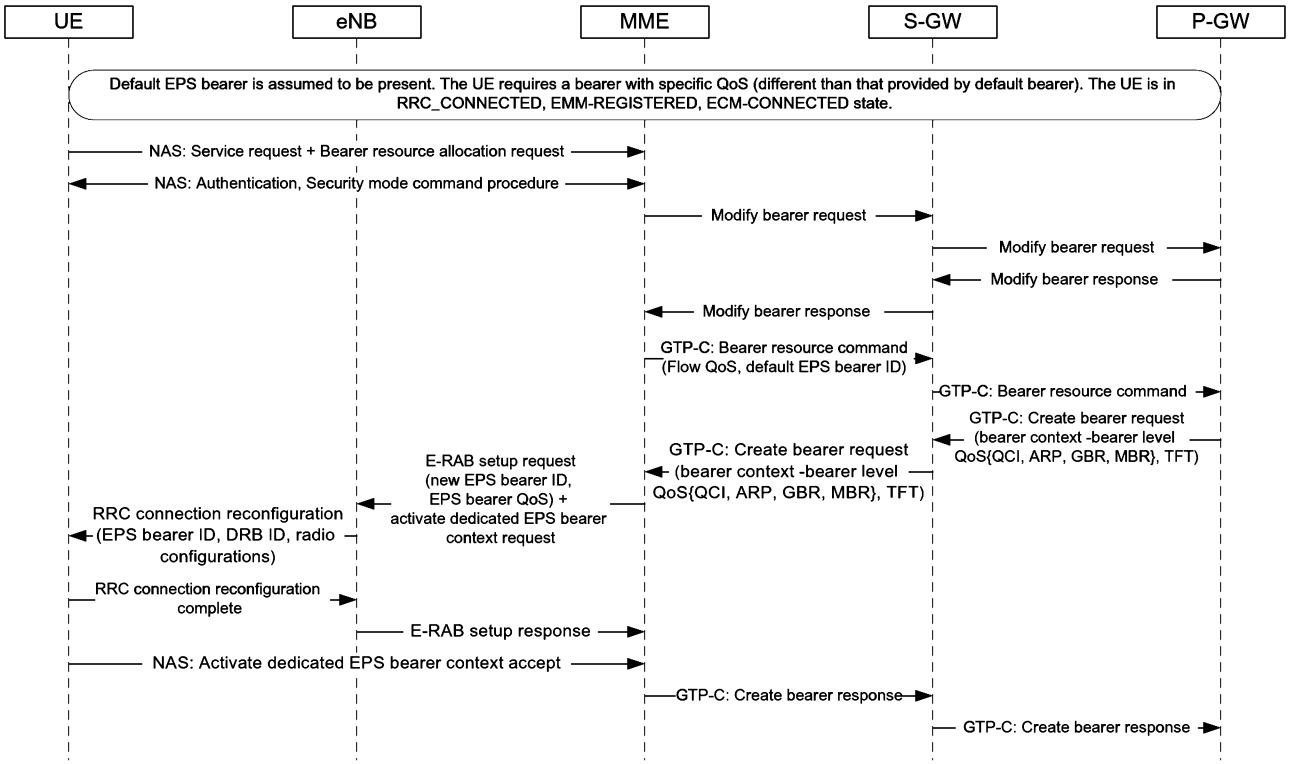


FIGURE 4.14

Call flow for mobile-originated data call [25].

receiving the modify bearer request, the S-GW activates the required resources and forwards the modify bearer request to the P-GW.

4. The P-GW processes the modify bearer request message and activates the required resources. Note that the IP address is allocated during the attach procedure. It responds to S-GW with the modify bearer response message and the S-GW forwards this message to the MME.
5. The MME now initiates the dedicated bearer establishment procedure by sending the eGTP-C bearer resource command to the S-GW. The S-GW processes the bearer resource command and forwards it to the P-GW.
6. The P-GW responds to the S-GW with a create bearer request message after allocating the dedicated bearer resource (TFT initialization). The S-GW processes the create bearer request and forwards it to the MME for further processing.
7. The MME now sends the E-RAB setup request to the eNB to set up the bearer between the eNB and the S-GW. It also piggybacks the NAS-activate-dedicated EPS-bearer context request to the UE.
8. The eNB allocates the resources for the radio bearers using an RRC connection reconfiguration request message to the UE and includes the received NAS message. The UE establishes the radio bearers and responds with an RRC connection reconfiguration complete message to the eNB.
9. At this time, the radio bearers are established between the eNB and the UE; thus the eNB sends the E-RAB setup response to the MME. The UE sends the activate dedicated EPS-bearer context accept NAS message to the MME via the eNB.
10. The MME sends a create bearer response to the S-GW to complete the dedicated bearer activation. The S-GW then forwards the message to the P-GW.

Once the UE concludes a data call, it can trigger the release of the dedicated bearers by sending the *Bearer Resource Modification Request* message to the MME, which can then take care of releasing the dedicated bearer with the S-GW and P-GW.

4.6.2 UE-terminated data call procedure

For a UE-terminated call, the network sends the paging request to all the eNBs associated with the last known tracking area where the UE was located. When receiving the *Paging Request* message from the MME, the eNB sends the paging message over the radio interface in the cells which are contained within one of the tracking areas provided in that message. The UE is normally paged using its S-TMSI or IMSI. The paging message also contains a UE ID index value in order for the eNB to calculate the paging occasions at which the UE listens to the paging message. The procedure is depicted in Figure 4.15, and described as follows [25]:

1. The P-GW/S-GW receives the incoming IP packet addressed to a UE.
2. The S-GW sends a downlink data notification to the MME requesting creation of a dedicated UE bearer.

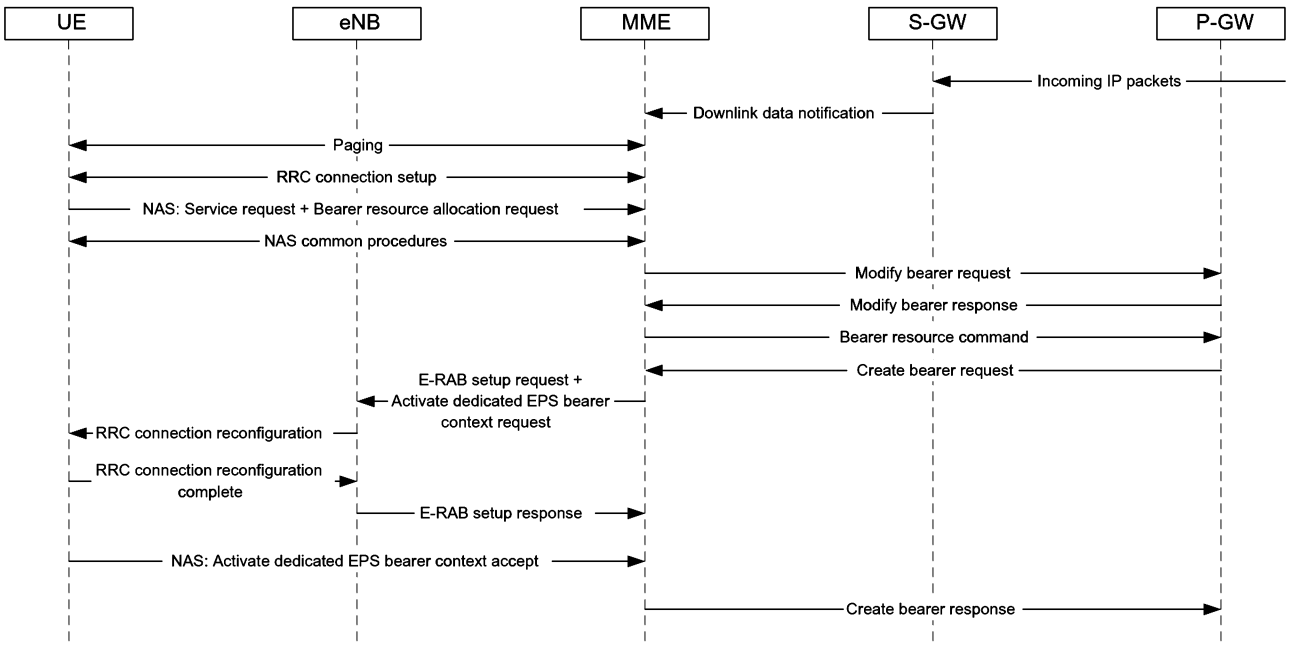


FIGURE 4.15

Call flow for mobile-terminated data call [25].

3. The MME now sends a paging message to notify the UE about the incoming IP packet.
4. Once the UE receives the paging message on the radio interface, it establishes an RRC connection with the eNB.
5. The UE sends the service request message to the MME and includes the bearer resource allocation request to request the dedicated bearer establishment. From this point onward, the message sequence is the same as that outlined in the mobile-originated data call.

4.7 Mobility procedures

E-UTRA supports X2-based and S1-based handover types. The X2-based handover is the least complex form of handover from an operation point of view, and typically would be used when there is an X2 interface between the source eNB and the target eNB. The handover is negotiated directly between the two eNBs; when the UE context is established on the target eNB, the MME is notified in order to switch the path. When no X2 interface exists between the network nodes, the S1-based handover will be utilized with signaling between the target and source eNBs going through the MME.

All handovers in E-UTRA are hard handovers, i.e., the air-interface connection with the source eNB should be disconnected before reestablishing a new connection with the target eNB [20]. The procedure for reestablishing the connection with the target node is identical for both X2-based and S1-based handovers, and requires the use of the random access procedure to synchronize and begin uplink transmissions. The UE is typically notified of the specific random access signature to use so that a non-contention-based procedure can be performed. From the core network perspective, handover can involve changes to the serving MME and S-GW, and this can be due to mobility and the geographic location of the network nodes, or due to loading and congestion control mechanisms. Figure 4.16 illustrates the overall concept of handovers and different handover types in E-UTRA. The X2-based handover procedure is performed without EPC involvement, i.e., preparation messages are directly exchanged between the eNBs. The release of the resources at the serving eNB (or source eNB) during the handover completion phase is triggered by the target eNB.

In order to support mobility, a UE may conduct and report to eNB a series of measurements at the physical layer which include RSRP and E-UTRA-carrier received signal strength indicator. These measurements in E-UTRA are classified as follows: intra-frequency E-UTRAN measurements; inter-frequency E-UTRAN measurements; and inter-RAT measurements. Measurement commands are used by E-UTRAN to instruct the UE to start measurements, modify measurements, or stop measurements. The reporting criteria that are used include event triggered reporting, periodic reporting, and event triggered

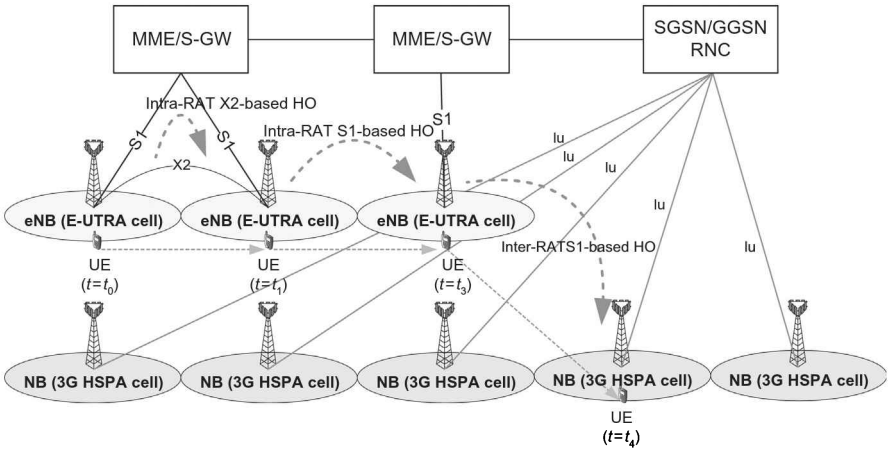


FIGURE 4.16

Overview of LTE handover scenarios [30].

periodic reporting. In the RRC_CONNECTED state, network-controlled UE-assisted handovers are performed and various DRX cycles are supported. In the RRC_IDLE state, cell reselections are performed and DRX is supported. The E-UTRAN can configure the UE to perform radio quality measurements for handover purposes. Different measurement objects are defined in the specification [9], and one object can be configured for each intra-frequency, inter-frequency, or inter-RAT measurement.

Depending on whether the UE needs transmission/reception gaps to perform the relevant measurements, measurements are classified as gap assisted or non-gap assisted. A non-gap-assisted measurement is a measurement on a cell that does not require transmission/reception gaps to allow the measurement to be performed. A gap-assisted measurement is a measurement on a cell that does require transmission/reception gaps to allow the measurement to be performed. Gap patterns are configured and activated by the RRC sublayer [3].

The measurements to be performed by a UE for intra/inter-frequency mobility can be controlled by E-UTRAN, using broadcast or dedicated control signaling. In the RRC_IDLE state, a UE follows the measurement parameters defined for cell reselection. In the RRC_CONNECTED state, a UE follows the measurement configurations specified by RRC sublayer and managed by the E-UTRAN. Intra-frequency and inter-frequency neighbor cell measurements are defined as follows [3]:

- Neighbor-cell measurements performed by the UE are intra-frequency measurements when the source and target cell operate on the same carrier frequency.

- Neighbor-cell measurements performed by the UE are inter-frequency measurements when the neighbor cell operates on a different carrier frequency compared to the source cell.

The mobility procedures in E-UTRA can be divided into idle mode and connected mode mobility for an "attached" UE. The idle mode mobility is based on UE autonomous cell reselections according to the parameters provided by the network. The UE transition between the idle and the connected mode is controlled by the network according to the UE activity and mobility.

Figure 4.17 illustrates the difference between the idle mode and connected mode mobility schemes. As shown in the figure, mobility in the RRC_IDLE state is UE-controlled using cell reselection, while in the RRC_CONNECTED state it is controlled by the E-UTRAN using handover. Furthermore, the network knowledge of the location of the UE in the RRC_IDLE state has a granularity of a tracking area, whereas the network knowledge of the UE location in the RRC_CONNECTED state has a granularity of a cell. A UE in the RRC_IDLE

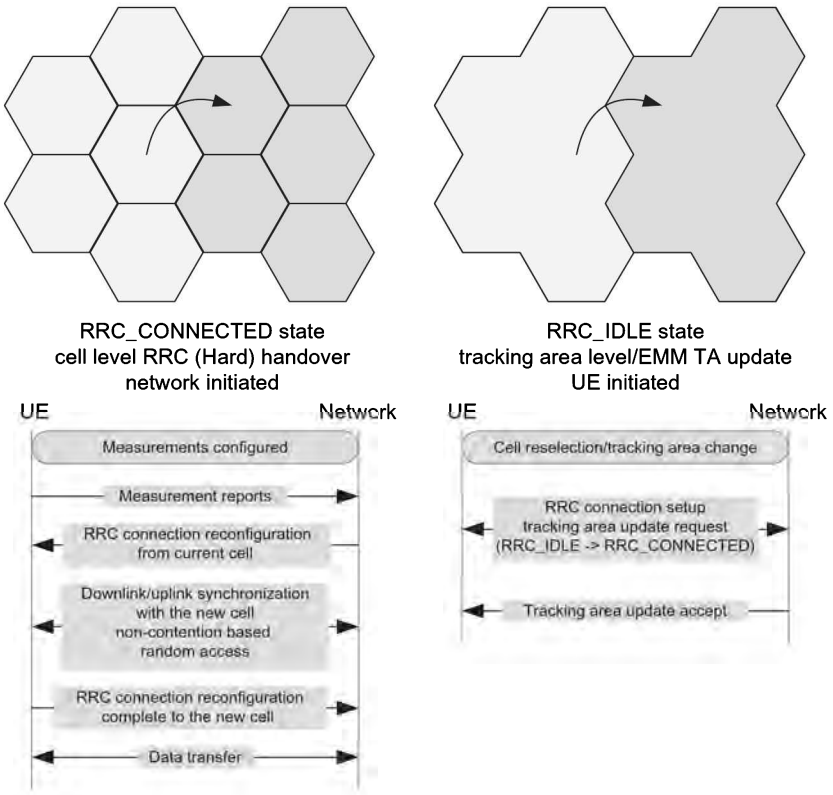


FIGURE 4.17

Overview of idle mode and connected mode mobility procedures [37].

state performs cell reselection and a tracking area update while crossing the tracking area boundaries. A UE in the RRC_CONNECTED state is configured to conduct radio link measurements and send them to the serving eNB. It can further receive/transmit data in the downlink/uplink and perform a non-contention-based random access procedure to stay uplink-synchronized with the network.

In the RRC_IDLE state, if equal priorities are assigned to multiple cells, the cells are ranked based on radio link quality. Equal priorities are not applicable between frequencies of different RATs. The UE does not consider frequencies for which it does not have an associated priority [20].

In RRC_CONNECTED, the E-UTRAN decides to which cell a UE should be transferred in order to maintain a reliable and satisfactory radio link quality. The E-UTRAN may take into account not only the radio link quality, but also other criteria such as UE capability, subscriber type, and access restrictions. Although E-UTRAN may trigger handover without measurement information from the UE, which is referred to as blind handover, typically it configures the UE to report measurements of the candidate target cells. Besides the handover procedure, the E-UTRA also allows a UE to be redirected to another frequency or RAT upon RRC connection release. Redirection during connection establishment is not supported, since at that time the E-UTRAN may not have all relevant information such as the capabilities of the UE and the type of subscriber [20].

4.7.1 Idle mode mobility procedures

In the RRC_IDLE state, cell reselection between frequencies is based on absolute priorities, where each frequency has an associated priority. Cell-specific default values of the priorities are broadcast via the system information. In addition, E-UTRAN may assign UE-specific values upon connection release, considering certain criteria such as UE capability or subscriber type. In case equal priorities are assigned to multiple cells, the cells are ranked based on radio link quality. Equal priorities are not applicable between frequencies of different RATs. The UE does not consider frequencies for which it does not have an associated priority. Besides cell reselection priority of a frequency, there are no idle mode mobility-related parameters that may be assigned via dedicated signaling.

In general, the transition from RRC_IDLE to RRC_CONNECTED is under the assumption that the RRC protection keys and user-plane protection keys are generated, while keys for NAS protection as well as higher-layer keys are assumed to be already available at the MME. On the other hand, in transitions from RRC_CONNECTED to RRC_IDLE, the eNBs delete the security keys that they store; thus the states of idle mode UEs are only stored in the MME.

In LTE-Advanced, the requirement for UE transition latency from the idle mode, with allocated IP address, to connected mode is less than 50 ms including the establishment of the user-plane and excluding the S1 transfer delay. The transition latency requirement from dormant mode (inactivity periods in short/long

DRX mode) to active mode (continuous transmission/reception periods) in connected mode is less than 10 ms. These requirements are stricter than those for earlier releases of LTE. In the dormant mode, the UE has an established RRC connection and radio bearers. The UE has already been identified at cell level, but it may be in DRX mode to save power during inactivity periods.

4.7.2 Connected mode mobility procedures

In the RRC_CONNECTED state, the network controls the UE mobility, i.e., the network decides when and where the UE should initiate a handover process. For network-controlled mobility in the RRC_CONNECTED state, the network triggers the handover process based on radio conditions or network load condition. To facilitate this process, the network may configure the UE to perform and report measurements including the configuration of measurement gaps. The network may also initiate a blind handover, i.e., without waiting for the measurement reports from the UE. Before sending the handover message to the UE, the source eNB prepares one or more target cells. The target eNB generates the message that is used to perform the handover, i.e., the message includes the AS configuration to be used in the target cell. The source eNB transparently (i.e., does not modify values or content) forwards the handover information received from the target cell to the UE. When appropriate, the source eNB may initiate data forwarding for (a subset of) DRBs in order to reduce the handover interruption time.

After receiving the handover message, the UE attempts to access the target cell at the first available random access channel opportunity according to random access resource selection, i.e., the handover is asynchronous. Consequently, when allocating a dedicated preamble for the random access in the target cell, E-UTRA ensures it is available from the first random access opportunity that the UE may use. Upon successful completion of the handover, the UE sends a message to confirm the handover. After successful completion of the handover, PDCP SDUs may be retransmitted in the target cell. This only applies to DRBs using RLC acknowledged mode [3,7]. The source eNB should, for some time, maintain a context to enable the UE to return in case of handover failure. In the case of handover failure, the UE attempts to resume the RRC connection either in the source or another cell using the RRC reestablishment procedure. This connection resumption succeeds only if the accessed cell is prepared, i.e., it concerns a cell of the source eNB or of another eNB for which handover preparation has been performed.

In LTE, the UE always connects to a single cell; thus the switching of a UE's connection from a source cell to a target cell is a hard handover. In a hard handover process, the eNB which controls the source cell requests the target eNB to prepare for the handover. The target eNB subsequently generates the RRC message in order to instruct the UE to perform the handover, and the message is transparently forwarded by the source eNB to the UE. In another type of handover

supported by LTE, the UE spontaneously decides to connect to the target cell, where it requests that the connection be continued. The UE applies this connection reestablishment procedure only after terminating the connection to the source cell. The latter procedure only succeeds if the target cell has been prepared in advance for the handover. In addition, LTE also allows a UE to be redirected to another frequency or RAT upon connection release (see Figure 4.1). Redirection during connection establishment is not supported, since at that time the E-UTRAN may not have all relevant information such as UE capability and the type of subscriber. However, the redirection may be performed while AS security has not been activated. When redirecting the UE to UTRAN or GERAN, the E-UTRAN may provide the system information for one or more cells on the relevant frequency. If the UE selects one of the cells for which the system information is provided, then it does not need to acquire it [3,20,21].

4.7.3 HetNet mobility

In general, the mobility issues are the same in heterogeneous networks and (homogeneous) macro-cell networks, i.e., managing inter-cell handovers in the connected mode and idle mode mobility (Figure 4.18). The presence of a small-cell layer and small-cell sizes creates a new opportunity and gives rise to additional challenges. The first issue is how to manage idle mode UE mobility between a small cell and a macro-cell. Camping strategy is an essential part of

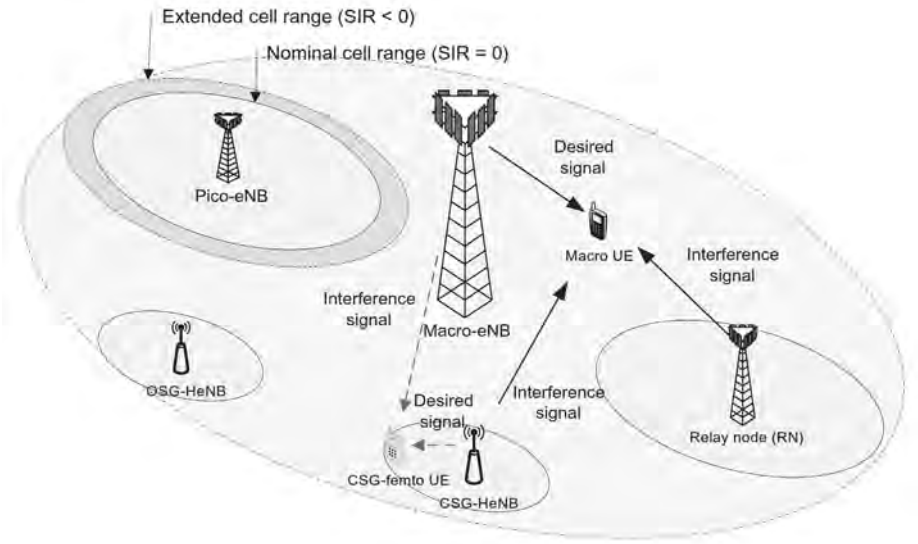


FIGURE 4.18

Illustration of a heterogeneous network and UE mobility.

idle mode cell reselection parameters as it affects the way UEs reselect and camp on macro-cells versus small cells. This fundamentally determines the idle mode UE distribution (cell loading) between macro-cells and small-cells, and affects relative loading between the two layers. The camping strategy also has a potential impact on UE power consumption as well as frequency of handovers. Increased number of handovers due to small-cell sizes creates a potential source of increased dropped call rate, even when handover success rate remains unchanged. A higher failure rate has been observed in the studies for outbound mobility from a small cell for high-speed UEs. This can be mitigated by having the higher-speed UEs confined to the macro-cell layer, but achieving this requires speed-based traffic management which is a challenge for LTE networks. In a limited number of scenarios, depending on camping strategy, operator preference, etc., connected mode UEs in the macro-layer may need to be handed over to a small cell. In such cases, if the small-cell operates on a different RF carrier, efficient small-cell discovery that minimizes impact on UE battery life is also necessary [38]. The number of radio link failures and handover events per UE per second for macro-only and HetNet scenarios is shown in Figure 4.19.

One of the observations which was made during heterogeneous networks mobility studies in 3GPP (results shown in Figures 4.19 and 4.20) suggests that among various possibilities for mobility in a heterogeneous network, the outbound mobility from a small cell has a relatively higher failure rate than other types of heterogeneous network mobility. In particular, the failure rate is higher for fast moving UEs [13]. While decisions regarding the need for enhancements and possible solutions are still under investigation, this is a practical case that a fast moving UE needs to be kept away from a small-cell layer in order to improve the robustness of heterogeneous network mobility.

Mobility management is performed to support UE mobility, which involves informing the network of UEs' present locations at the cell ID (or alternatively HeNB) level, providing the network with UE IDs, and maintaining physical channels. As discussed earlier, LTE mobility management is classified into two states depending on the UE's corresponding RRC connection state, i.e., idle mode and connected mode mobility management. For each of the states, intra-frequency, inter-frequency, or inter-RAT mobility procedures can be performed based on the UE measurements of available frequencies/RATs.

In connected mode mobility management, the mobile network ensures continuity of transmission of physical signals/channels and provides uninterrupted communication services to UEs through handover when the UEs move in the network. Handover is a procedure where the serving cell of a UE in the connected mode is changed. In LTE, the handover decisions are performed by the eNBs without involving the MME or serving gateway, which is used as an anchor node. That is, the eNB delivers the associated configuration through signaling on the control-plane, UEs perform measurements accordingly and complete the handover procedures under the control of the eNB. A UE that is powered on but does not have an RRC connection to the access network is in

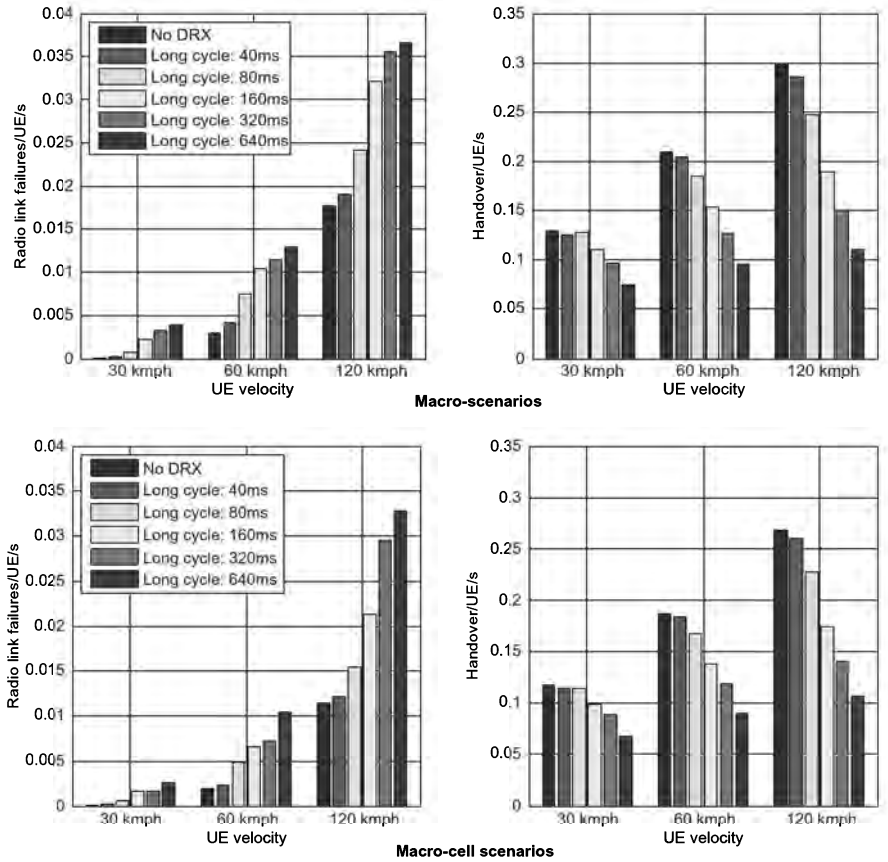


FIGURE 4.19

Comparison of radio link failure and handover events in macro- and HetNet scenarios with various DRX configurations [13].

the idle mode. In this mode, UEs report their locations to the network and the eNB broadcasts system information, which can be used by the UEs to select suitable cells to camp on.

Small cells can be introduced to overlay the existing LTE macro-cell networks either by sharing the same RF carrier or using a dedicated RF carrier. If the existing macro-node RF carrier is shared by small cells, interference management techniques must be used to avoid performance degradation due to inter-cell interference. A notable exception to this rule is when the same-carrier small cell is deployed to overlay an existing macro-cell coverage hole. In this case, the two cells will be sufficiently isolated from the RF perspective. If small cells are deployed on a dedicated LTE carrier, neighbor relationship(s) must be established

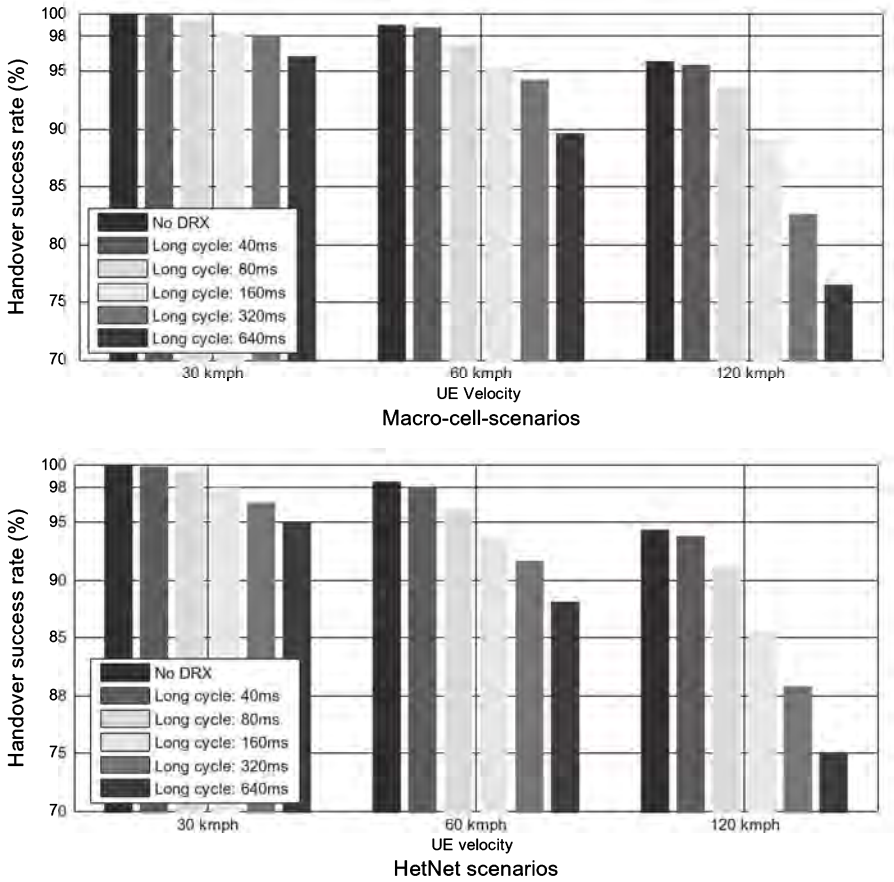


FIGURE 4.20

Comparison of handover success rate in macro-cell and HetNet scenarios with various DRX configurations [13].

with the existing LTE macro-cells (by advertising small-cell frequency in SIB5) to enable UE's in the macro-cell network to search for the small cells.

There are a number of options for establishing neighbor relationship with the macro-cell RF carriers. One option is that all macro-cell carriers list small-cell carrier as neighbor in SIB5. Another option is to allow only a fraction of macro-cell carriers to list small-cell carriers as neighbor in SIB5. For cells on the remaining RF carriers, inter-macro-carrier load-balancing techniques can be used to redistribute the load.

Depending on the option chosen, the camping and load balancing between the small cell and macro-carriers can be impacted. The first option provides immediate offloading to all macro-carriers in a small-cell coverage area, but the risk of

small cells becoming overloaded increases. The second option only provides immediate capacity relief to neighbor macro-cells.

During the cell selection procedure, the UE determines the best cell that can be used to establish an RRC connection. The UE will typically select a cell based on the RSRP that is calculated from measurements performed on the cell-specific reference signals broadcast by the cell with a nominal transmit power, i.e., Cell $ID_{serving\text{-}cell} = \arg\ \max_k\{RSRP(k)\}$. Because of the substantially higher power emitted by the macro-cell, the downlink coverage of the small cell is reduced. This issue can further create asymmetry in the uplink coverage that is unaffected by the transmit power differences of the cells. LTE Rel-8 allows addition of a bias to the cell selection with the UE determining the serving cell based on the following criterion: Cell $ID_{serving\text{-}cell} = \arg\ \max_k\{RSRP(k) + \Delta_{bias}(k)\}$[38]. The cell selection bias can range from 0 to 20 dB and can be communicated to the UE via higher layer signaling or can be cell specific, which in this case is broadcast as part of system information. The effect of the bias is to extend the downlink small-cell coverage as the UE artificially increases the measured RSRP from the small cell by the bias value (see Figure 4.18). At the same time, the downlink signal-to-interference plus noise ratio, when an RRC connection establishment is attempted is degraded. In cases where the small cell is sufficiently far from the macro-cell, the range of the small cell would be adequate without the bias, especially if the transmit power of the small cell is increased. In the latter case, increase of the transmit power is to some extent another method to increase the range of pico-cells.

For UEs in the connected state that are moving out of cell coverage, handovers are required to move their radio link from one cell to another. It is preferable to keep a UE in the same layer (small cell or macro-cell) as it performs handovers, unless there is a necessity to move it from one layer to the other. Same layer handovers use the normal coverage-based intra-frequency handover framework used in a macro-cell network.

In summary, the UE initiates a handover procedure, based on signal and time thresholds defined and transmitted to it by the serving cell. These are the handover hysteresis, time to trigger, and the cell individual offset. The latter is a vector quantity sent by the serving cell that contains the offsets of the serving and neighbor cells that all UEs in this cell must use in determining whether the measured RSRP satisfies the conditions of handover event A3 that results in the UE sending a measurement report that triggers the handover [9]. The individual offset values can be positive or negative. The measured RSRPs are filtered and individual offset values are added according to the following criterion: Cell $ID_{serving\text{-}cell} = \arg\ \max_k\{RSRP(k) + \Delta_{individual\text{-}offset}(k)\}$[38]. The challenge is to optimally configure the cells with these parameters such that the handover success rate and the dropped call rate meet the target values. Load distribution depends, among other factors, on the UE average mobility condition, potentially estimated from MME layer events and usually applied to disable ping-pong effects [38]. The time that a UE stays connected with a cell following a handover is used as the metric to evaluate the ping-pong behavior. The time-of-stay in cell A is the time

interval from the moment when the UE successfully sends a handover complete message, i.e., *RRC Connection Reconfiguration Complete* message, to cell A, until the UE successfully sends a handover complete message to cell B. The minimum time of stay connected with a cell models the time needed to allow a UE to establish a reliable connection with the cell, plus the time required for efficient data transmission. If a UE performs a handover from cell B to cell A, and then makes a handover back from cell A to cell B and the time connected to cell A was less than the minimum time-of-stay, it is considered as a ping-pong effect. In general, if the UE's time-of-stay in a cell is less than the minimum time-of-stay, the handover may be considered as unnecessary, resulting in unnecessary use of network resources [13].

4.7.4 S1-based handover procedure

The E-UTRAN supports mobility within the LTE system, as well as mobility to/from other systems using both 3GPP and non-3GPP radio access technologies [11]. The mobility procedures also involve the network interfaces. This section discusses the procedures over the S1 interface to support mobility. In this case, the UE will assist in the handover decision by measuring the neighbor cells and reporting the measurements to the network, which in turn decides upon the handover timing and the target cell/node. The parameters to measure and the thresholds for reporting are determined by the network.

The E-UTRA supports two types of handover for UEs in the connected mode: S1-based and X2-based handover (Figure 4.21). The X2-based handover

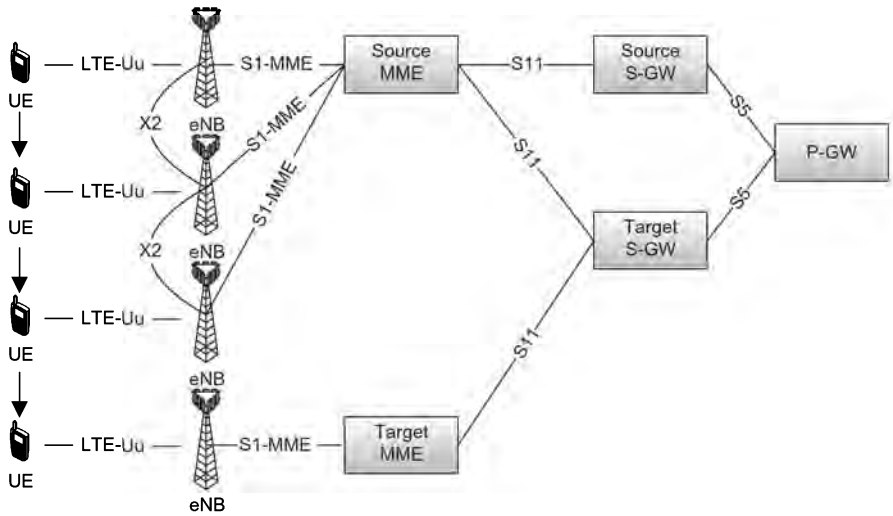


FIGURE 4.21

Illustration of an S1-based mobility scenario [23].

procedure is typically used for inter-eNB handover. However, when there is no X2 interface between the two eNBs, or if the source eNB has been configured to initiate handover to a particular target eNB through the S1 interface, then an S1-based handover will be triggered. The S1-based handover procedure, shown in Figure 4.22, consists of a preparation phase, where the core network resources are prepared at the target node, followed by an execution phase, and a completion phase. The *Status Transfer* message sent from the source eNB carries some PDCP status information that is needed at the target eNB in scenarios where PDCP status preservation applies for the S1-based handover [8]. This is consistent with the information that is sent within the X2 status transfer message used for the X2-based handover. As a result, handling of the handover by the target eNB as seen from the UE is exactly the same, regardless of the type of handover (S1 or X2). The status transfer procedure is assumed to be triggered concurrently with the start of data forwarding after the source eNB has received the *Handover Command* message from the source MME.

In an inter-MME handover, two MMEs (the source and target MME) are involved, as shown in Figure 4.21. The source MME controls the source eNB and the target MME manages the target eNB. In this case, both MMEs are connected to the same S-GW. This handover is triggered when the UE moves from one MME area to another MME area. The data forwarding can be either direct or indirect, depending on the availability of a direct path for the user-plane data between the source eNB and the target eNB. The *Handover Notify* message, which is sent later by the target eNB upon confirmation of the arrival of the UE at the target node, is forwarded by the MME to trigger path update in the S-GW to the target eNB. In contrast to the X2-based handover, the message is not acknowledged and the resources at the source eNB side are released later upon reception of a *Release Resource* message triggered by the source MME [23,28] (Figures 4.23 and 4.24).

Due to lack of direct connection between the source and target eNBs, the MME has to be involved in the process. The S1-based handover procedure is used when the X2-based handover cannot be used. The source eNB initiates a handover by sending a *Handover Required* message over the S1-MME reference point. This procedure may relocate the MME and/or the S-GW. The source MME selects the target MME. The MME should not be relocated[8] during inter-eNB handover unless the UE leaves the MME pool area where the UE is served. The MME (target MME for MME relocation) determines if the S-GW needs to be relocated. If the S-GW needs to be relocated, the MME selects the target S-GW. The source eNB decides which data packets corresponding to the downlink/uplink EPS bearers are going to be forwarded to the target eNB. Packet forwarding can take place either directly from the source eNB to the target eNB, or indirectly from the source eNB to the target eNB via the source and target S-GWs, or if the S-GW is not relocated, only through a single S-GW. The availability of a direct

[8]Relocation is defined as switching of the communication equipment such as area switches during communication.

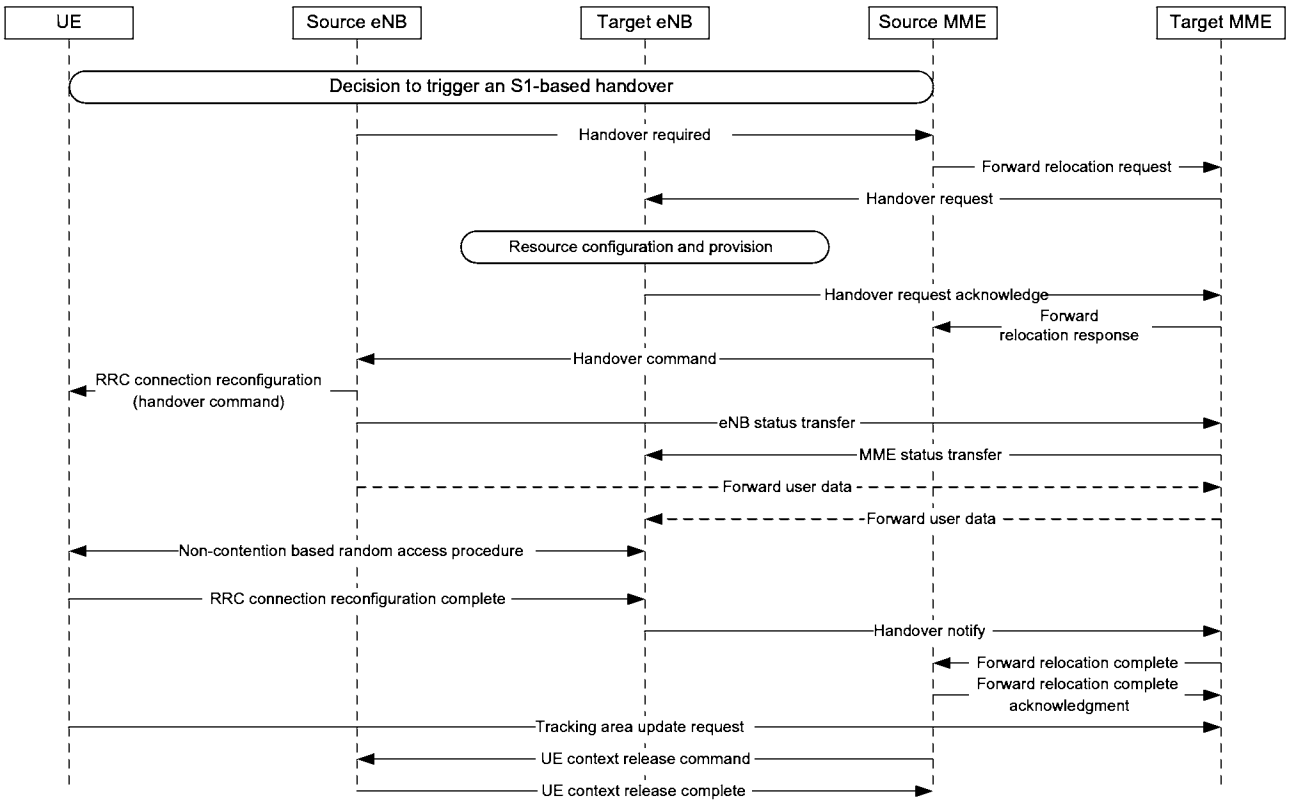


FIGURE 4.22

S1-based handover procedure [26,28].

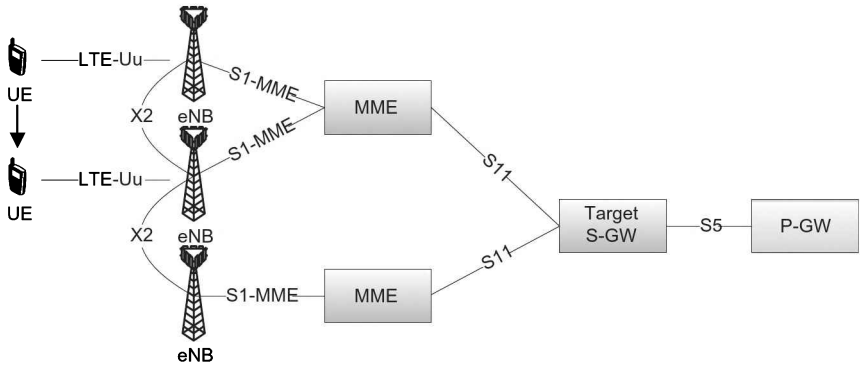


FIGURE 4.23

Illustration of X2-based handover with no S-GW change [23].

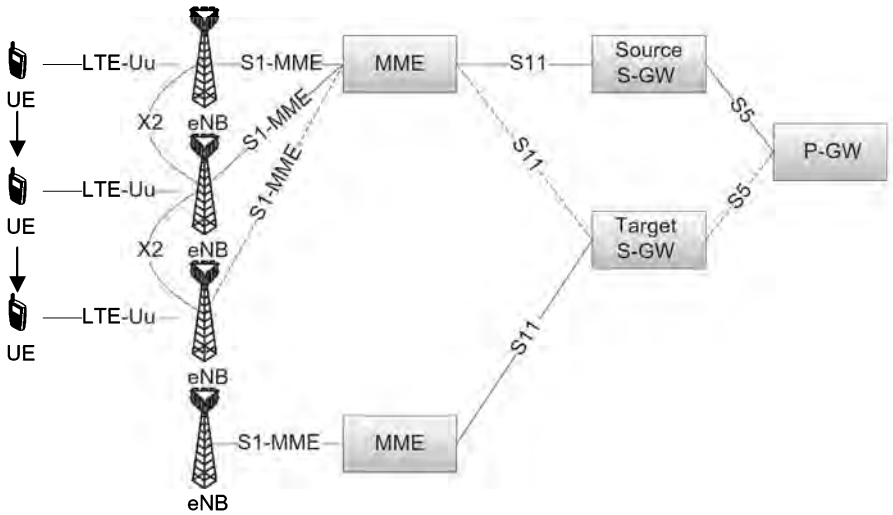


FIGURE 4.24

Illustration of an X2-based handover with change of S-GW [23].

forwarding path is determined in the source eNB and is indicated to the source MME. If X2 connectivity is available between the source and target eNBs, a direct forwarding path is typically available. If a direct forwarding path is not available, indirect forwarding may be used. The source MME uses the information from the source eNB to determine whether to apply indirect forwarding. The source MME informs the target MME whether indirect forwarding should be utilized.

Samsung Ex. 1005, Page 242 of 1147

4.7.5 X2-based handover without S-GW change

This procedure is used to hand over a UE from a source eNB to a target eNB using the X2 interface when the MME and S-GW are unchanged, which would be possible only if there is direct connectivity between the source and target eNBs with the X2 interface. The X2 handover procedure is performed without MME involvement, which implies that preparation messages are directly exchanged between the source and the target eNBs. The release of the resources at the source side during the handover completion phase is triggered by the target eNB. The call flow for this type of handover is shown in Figure 4.25. This procedure can be summarized as follows [26,28]:

1. Based on the measurement reports from the UE, the source eNB determines that the UE should be handed over to a target eNB using the handover procedures. Note that each network operator's criteria for handover might be different from other operators. The source eNB may send a *Resource Status Request* message to determine the loading condition at the target eNB. Based on the received *Resource Status Response* message, the source eNB can make a decision to proceed with the handover procedure using the X2 interface.
2. The source eNB sends a *Handover Request* message to the target eNB, transferring the necessary information such as UE context, which includes the security context and radio bearer context to prepare the handover at the target side. The target eNB checks the resource availability and reserves the resources and sends back the *Handover Request Acknowledgement* message including a transparent container that is sent to the UE as an RRC message to perform the handover. The container may include a new C-RNTI, target eNB

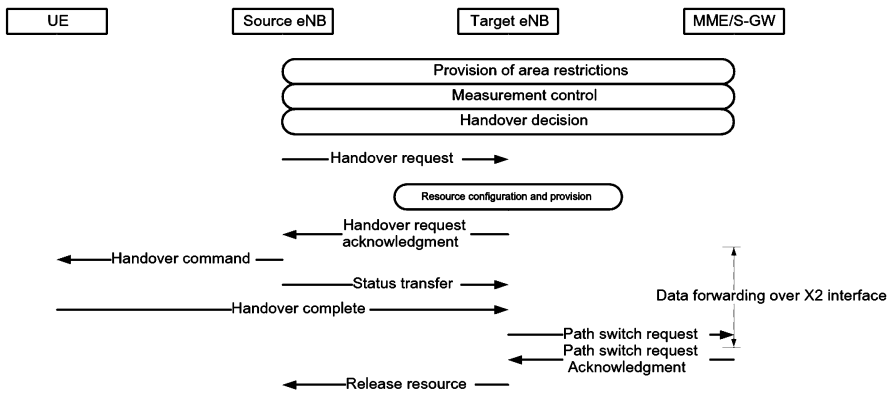


FIGURE 4.25

X2-based handover procedure [28].

security algorithm identifiers for the selected security algorithms, a dedicated RACH preamble, and access parameters.

3. The source eNB generates an *RRC Connection Reconfiguration* message to perform the handover including the mobility control information. The source eNB performs the necessary integrity protection and ciphering of the message and sends it to the UE. The source eNB further sends the *Status Transfer* message to the target eNB to convey the PDCP and HFN status of the E-RABs. For those bearers requiring in-sequence delivery of packets, the *Status Transfer* message provides the sequence number and the HFN that the target eNB should assign to the first packet with no sequence number. The first packet can either be the one received over the target S1 path or the packet received over X2, if data forwarding over X2 is used.
4. The source eNB starts forwarding the downlink data packets to the target eNB for all data bearers (which are being established in the target eNB during the *Handover Request* message processing). In the meantime, the UE tries to access the target eNB using the non-contention-based random access procedure. If it succeeds in accessing the target cell, it sends an *RRC Connection Reconfiguration Complete* to the target eNB.
5. The target eNB sends a *Path Switch Request* message to the MME to inform that the UE has changed cells, including the TAI + E-UTRAN cell global identifier of the target. The MME determines that the S-GW can continue to serve the UE [31].
6. The MME sends a *Modify Bearer Request* message to the S-GW, including the eNB address and TEIDs for downlink user-plane for the accepted EPS bearers. If the P-GW requested the UE's location information, the MME also includes the *User Location Information* in this message.
7. The S-GW sends the downlink packets to the target eNB using the newly received addresses and TEIDs (path switched in the downlink to the target eNB), and the *Modify Bearer Response* to the MME. The S-GW sends one or more end marker packets on the old path to the source eNB and then can release any user-plane/transport network layer resources toward the source eNB.
8. The MME responds to the target eNB with a *Path Switch Request Acknowledge* message to notify the completion of the handover. The target eNB now requests the source eNB to release the resources using the *UE Context Release* message, completing the handover procedure. The target eNB requests the source eNB to release the resources using the *UE Context Release* message.

These procedures are used to hand over a UE from a source eNB to a target eNB using the X2 interface. In these procedures, the MME stays unchanged. Two procedures are defined depending on whether the S-GW is changed. In addition

to the X2 interface between the source and target eNB, the procedures rely on the presence of an S1-MME interface between the MME and the source eNB, as well as between the MME and the target eNB.

If during an X2-based handover the serving PLMN changes, the source eNB informs the target eNB of the PLMN selected to be the new PLMN. When the UE receives the handover command, it will remove any EPS bearers for which it did not receive the corresponding EPS radio bearers in the target cell. As part of handover execution, downlink/uplink packets are forwarded from the source eNB to the target eNB. When the UE has arrived in the target eNB area, downlink data forwarded from the source eNB can be sent to it. Uplink data from the UE can be delivered via the (source) S-GW to the P-GW, or optionally forwarded from the source eNB to the target eNB. Only the handover completion phase is affected by a potential change of the S-GW, the handover preparation and execution phases are identical.

4.7.6 X2-based handover with S-GW change

Handover through the X2 interface is triggered by default for intra-LTE mobility unless there is no X2 interface established or the source eNB is configured to use S1-based handover. The X2-based handover procedure is illustrated in Figure 4.25. Similar to an S1-based handover, the procedure is composed of a preparation phase, an execution phase, and a completion phase. The key features of an X2-based handover include direct transfer between the source and target eNBs (facilitating preparation phase); data forwarding per bearer in order to minimize packet loss, where the MME is only informed at the end of the handover procedure when the handover is successfully completed in order to trigger the path switch; and direct release of resources at the source side triggered by the target eNB. For those bearers requiring in-sequence delivery of packets, the *Status Transfer* message provides the sequence number and the HFN that the target eNB should assign to the first packet with no sequence number that it must deliver. The first packet can either be the one received over the target S1 path or received over X2, if data forwarding over X2 is used. When the source eNB sends the *Status Transfer* message, it pauses its packet transmit/receive status update, i.e., it stops assigning PDCP sequence numbers to downlink packets and stops delivering uplink packets to the network. The mobility over X2 can be categorized according to its resilience to packet loss. A handover is seamless, if it minimizes the interruption time during the inter-node transfer of a UE, or the handover can be lossless, if no packets are lost. These handover types rely on data forwarding of user-plane downlink packets. The source eNB may decide to operate in one of these modes on a per EPS-bearer basis, based on the QoS received over S1 for this bearer and the service characteristics.

If the source eNB selects the seamless handover mode for one bearer, it informs the target eNB in the *Handover Request* message to establish a

GTP tunnel[9] to enable downlink data forwarding. The target eNB may send an acceptance in the *Handover Request Acknowledgement* message at the tunnel endpoint where the forwarded data is expected to be received. The tunnel endpoint may be different from the one that has been set up as the termination point of the new bearer which is established over the target S1 [28].

Upon receipt of the *Handover Request Acknowledgement* message, the source eNB can start forwarding the arriving data over the source S1 path toward the indicated tunnel endpoint, concurrent with the transmission of the handover trigger to the UE over the radio interface. The forwarded data is going to be available at the target eNB for early delivery to the UE. When forwarding is in use and in-sequence delivery of packets is required, the target eNB is assumed to first deliver the packets forwarded over X2 before delivering the packets received over the target S1 path, once the S1 path switch has been completed. The end of the forwarding is signaled over X2 to the target eNB prior to switching the S1 path [28].

If the source eNB selects the lossless handover mode for one bearer, it will additionally forward (over X2) those user-plane downlink packets that are processed through the PDCP sublayer, but they are still locally buffered because they have not been delivered and acknowledged by the UE. Those packets are forwarded together with their assigned PDCP sequence number included in a GTP extension header field. They are sent over X2 prior to the arriving packets from the previous S1 path. The same mechanisms described above for seamless handover are used for the GTP tunnel establishment. The end of forwarding is also managed in the same way, since in-sequence packet delivery applies to lossless handovers, as well. The target eNB must ensure that all packets, including the ones received with a sequence number over X2, are delivered in sequence to the target node.

When lossless handover is used, the target eNB may not deliver some of the forwarded downlink packets received via X2 over the radio interface, if it is informed by the UE that those packets have already been received via the source eNB. This new feature is known as downlink selective retransmission in E-UTRA literature [20,28]. In the uplink, the target eNB may instruct the UE not to retransmit packets already received from the source eNB. In order to use uplink selective retransmission, the source eNB forwards the user-plane uplink packets that are received out of sequence to the target eNB via a new GTP tunnel. At the request of the target eNB, the source eNB establishes a new forwarding tunnel. In this

[9]GTP-U protocol is used over S1-U, X2, S4, S5, and S8 interfaces of the evolved packet system. GTP-U tunnels are used to carry encapsulated PDUs and signaling messages between a given pair of GTP-U tunnel endpoints. The tunnel endpoint identifier (TEID) which is present in the GTP header indicates which tunnel a particular PDU belongs to. The transport bearer is identified by the GTP-U TEID and the IP address (source TEID, destination TEID, source IP address, destination IP address) [2,31].

Samsung Ex. 1005, Page 246 of 1147

case, the target eNB includes a GTP tunnel endpoint where it expects the forwarded uplink packets to be received in the *Handover Request Acknowledgement* message. The source eNB then indicates in the *Status Transfer* message for this bearer, the list of sequence numbers corresponding to the packets to be forwarded. Based on this information, the target eNB informs the UE of the packets that are not going to be retransmitted, making the overall retransmission scheme more efficient [20,28].

4.7.7 Inter-RAT mobility procedures

In principle, the inter-RAT handover procedures when an LTE system is the source or target are the same as the intra-LTE handover process per se. One of the differences is that upon handover to an LTE, the entire security association configuration needs to be signaled, whereas within an LTE it is possible to signal incremental information, where only the changes to the configuration are signaled. If ciphering was not activated in the previous RAT, the E-UTRAN activates ciphering as part of the handover procedure. The E-UTRAN also establishes SRB1, SRB2, and one or more DRBs, i.e., at least one DRB is associated with the default EPS bearer. Figure 4.26 illustrates the UE states in E-UTRA, UMTS, and GSM, as well as the mobility support between E-UTRAN, UTRAN, and

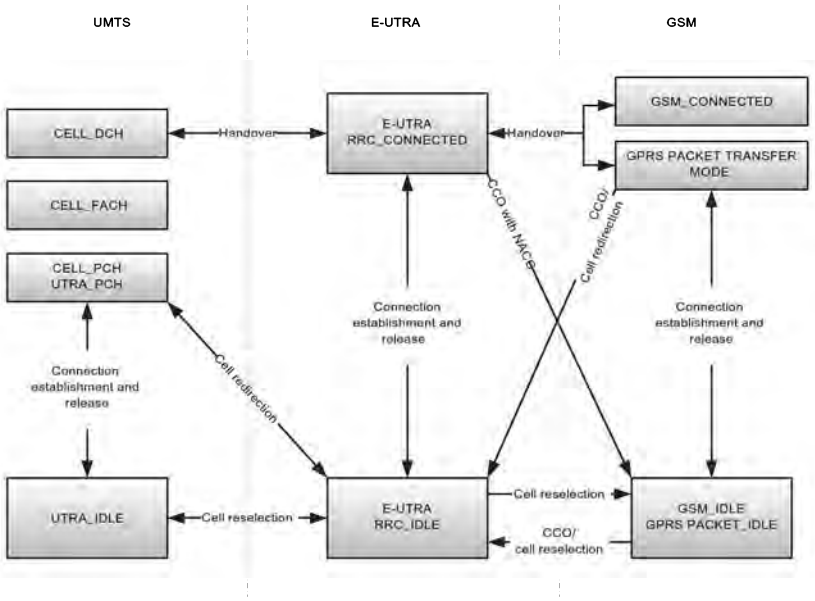


FIGURE 4.26

E-UTRA states and inter-RAT mobility procedures [3].

GERAN [10]. The cell change order (CCO) message is used to command the UE to change to another RAT [3].

In the context of multi-RAT operation, the terms cell selection, cell reselection, cell redirection, handover, and circuit-switched fallback (CSFB) are frequently used. In order to avoid confusion when using these terms, and to better understand the use cases, we use the illustration shown in Figure 4.27 to underline the significance and implication of each term. While this figure may oversimplify the process, it can give the reader some intuitive understanding about multi-cell, multi-RAT operation of E-UTRA and interaction with other radio access technologies. As shown in Figure 4.27, when a UE is powered on, it performs a cell search to find and select a suitable cell to camp on. The cell reselection, on the other hand, is an idle mode procedure and a mechanism to change cell after a UE is camped on a cell. Handover is a connected mode procedure and is performed between connected states of the same or different radio access technologies, e.g., E-UTRA RRC_CONNECTED state to UMTS CELL_DCH state. Redirection is a process which changes the UE state from connected to idle and in many cases for CSFB purposes, e.g., the UE is redirected from the E-UTRA RRC_CONNECTED state to the UMTS/GSM idle mode. In the case of CSFB, a handover is not required and possibly not supported by the network and/or the UE. In such a case, redirection is the default mechanism to utilize. Furthermore, in some cases, the redirection can be faster than handover due to reasons such as inter-RAT

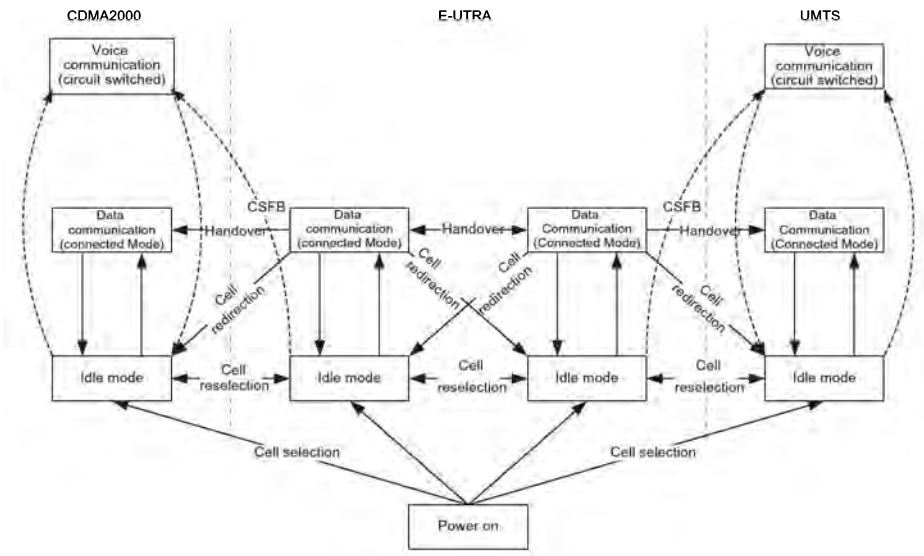


FIGURE 4.27

Illustration of cell selection, cell reselection, cell redirection, circuit-switched fallback, and handover in E-UTRA [32].

measurement delay/cell detection, layer-3 filtering,[10] time-to-trigger, handover preparation in the network, especially in cases where mobile originated or terminated the call, which requires the UE to transition to the RRC_CONNECTED state and to perform CSFB is triggered from the E-UTRA idle mode.

The CSFB functionality enables the support of voice service without deployment of IMS. This is a possible scenario for many operators when LTE is initially deployed because those operators may not support IMS in their networks. The CSFB functionality and IMS can be deployed simultaneously, meaning that an operator can gradually roll out an IMS system while still supporting a fallback mechanism whenever necessary. The UE informs the network that it wishes to perform a "Combined Attached" during initial registration with the MME. In practice, this means that the device requests the network to register its presence in UMTS/GSM circuit-switched networks, as well. When a mobile-terminated call for a subscriber (with multi-RAT support) arrives at the mobile switching center, the incoming call is signaled to the MME. The UE is paged in the LTE system, if it is in the idle mode, or is notified of the call, if it is in the connected mode. It responds by requesting CSFB for the call to proceed. For mobile-originated calls, the UE establishes an RRC connection, if in the idle mode and then notifies the MME that CSFB support for the call is required.

The mobility between E-UTRAN, CDMA2000 1xRTT [15,16], and CDMA2000 high rate packet data (HRPD) or CDMA 1xEV-DO [14] radio access technologies is shown in Figure 4.28. In general, the procedure for handover from LTE to another RAT supports both handover and CCO with network assisted cell change (NACC). The CCO or NACC procedure is applicable only for handover to GERAN. Handover from LTE is performed only after security has been activated. In the case of CSFB to CDMA2000, the procedure includes support for parallel handover (i.e., to 1xRTT and HRPD), for handover to 1xRTT combined with redirection to HRPD, and for redirection to HRPD (see Figure 4.28). The UE may assist by sending a *Measurement Report* message. In the case of handover (as opposed to CCO), the source eNB requests the target node to prepare for the handover in advance. As part of the handover preparation request, the source eNB provides information about the applicable inter-RAT UE capabilities, as well as information about the currently established bearers to the target node. In response, the target node generates the handover command and returns this to the source eNB. The source eNB sends mobility from the E-UTRA command to the

[10]The UE is required to apply a layer 3 filtering to each measurement quantity that it performs, which does not include quantities configured exclusively for UE receive–transmit time difference measurements. The UE is required to filter the measured result, before using for evaluation of reporting criteria or for measurement reporting, by the following formula $F_n = (1 - a)F_{n-1} + aM_n$, where M_n is the latest received measurement result from the physical layer, F_n is the updated filtered measurement result, that is used for evaluation of reporting criteria or for measurement reporting, and F_{n-1} is the previous filtered measurement result. The value of F_0 is set to M_1 when the first measurement result from the physical layer is received. The parameter a is defined as $a = 1/2^{k/4}$, in which k is the *filter coefficent* for the corresponding measurement quantity [9].

Samsung Ex. 1005, Page 249 of 1147

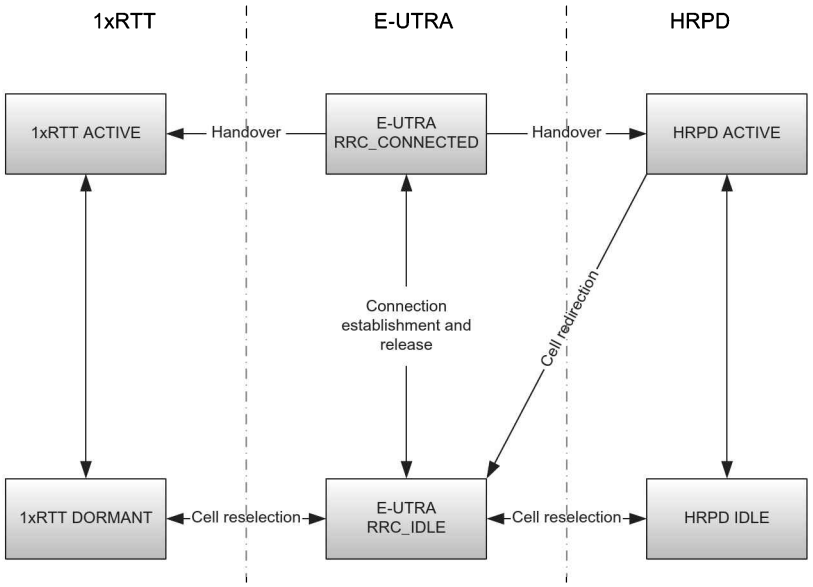


FIGURE 4.28

Mobility procedures between E-UTRA and CDMA2000 1xRTT/HRPD [3].

UE, which includes the inter-RAT message received from the target cell and some inter-RAT parameters in the case of CCO. Upon receiving the message, the UE starts the timer T304 and connects to the target node, either using the received radio configuration (handover) or by initiating connection establishment (CCO) according to the applicable specifications of the target RAT. The upper layers in the UE are informed by the AS of the bearers established by the target RAT base station. As a result, the UE is informed whether some of the previously established bearers were not admitted by the target RAT base station [3,20].

Having discussed inter-RAT operation, let us take LTE-to-UMTS inter-RAT handover as an example and describe the procedure involved in this operation. In this scenario, the source eNB connects to the source MME and source S-GW while the target UMTS radio network controller (RNC) connects to the target SGSN and target S-GW; however, both source and target S-GWs connect to the same P-GW. As shown in Figure 4.29, the source eNB configures the UMTS measurements in the UE using the *RRC Connection Reconfiguration* message. This message sets the thresholds (serving cell and inter-RAT neighbor cell) for activation, and further sets the events and provides UMTS center frequency and the scrambling code to be used. The UE measures the neighboring UMTS cells and sends RRC measurement reports back to the source eNB. The source eNB then makes a decision to perform a handover to the UMTS. In order to prepare the target UMTS cell, the source eNB sends a *Handover Required* message to the

Samsung Ex. 1005, Page 250 of 1147

source MME. This message contains the target UMTS cell identifier. The source MME then contacts the target SGSN using the UMTS cell-identifier and allocates the resources using a forward relocation request. Once the source MME receives confirmation (*Relocation Complete* message from target SGSN), the source MME sends a *Handover Command* message to the source eNB. The source eNB then sends an RRC *Mobility from E-UTRAN Command* to the UE, indicating the move to UTRAN. The UE then performs UTRAN access procedures to synchronize and attach to the target RNC and upon successful completion of the process, the UE sends a *Handover to UTRAN Complete* message to the target RNC [20,28,30]. Note that when the source eNB receives the handover command from the source MME, it begins forwarding data packets received from the S-GW. Following this process, downlink packets for the terminal that arrive at the S-GW are forwarded to the terminal through S-GW → source eNB → S-GW → target SGSN → target RNC (Figure 4.29).

Following the above procedure, the terminal switches to 3G, and when the radio link configuration is completed, notification that it has connected to the 3G radio access system is sent over each of the links to the source MME. The source MME sends a forward relocation complete acknowledgment to the target SGSN. A given period of time after receiving this message, the target SGSN releases the resources related to data forwarding. The target SGSN then sends a modify bearer request to the S-GW to switch the path between the S-GW and eNB to that between the S-GW and the target SGSN. This message contains information elements required to configure the path from the S-GW to the target SGSN, including those issued by the target SGSN. When the S-GW receives this signal, it configures a communications path from the S-GW to the target SGSN. In this way, the new communication path becomes S-GW → target SGSN → target RNC → terminal, and data transmission to the target 3G radio access system begins. Note that after this point, data forwarding is no longer needed, thus the S-GW sends a packet to the eNB with an "end marker" attached, and when the eNB receives this packet, it releases its resources related to data forwarding. The S-GW sends a modify bearer response to the target SGSN, indicating that the handover procedure has completed. The MME also releases eNB resources that are no longer needed [29].

In the above inter-RAT handover procedure, we described the handover procedure between 3GPP radio access systems in which the S-GW does not change; however, handovers with S-GW relocation are also possible. In those cases, the P-GW provides the anchor function[11] for path switching similar to switching to non-3GPP access systems.

In a UMTS-to-LTE inter-RAT handover, the source RNC connects to the source SGSN and source S-GW, the target eNB connects to the target MME and target S-GW, and both source and target S-GWs connect to the same P-GW.

[11]Anchor function is a function which switches the communication path according to the area where the terminal is located and forwards packets for the terminal to that area.

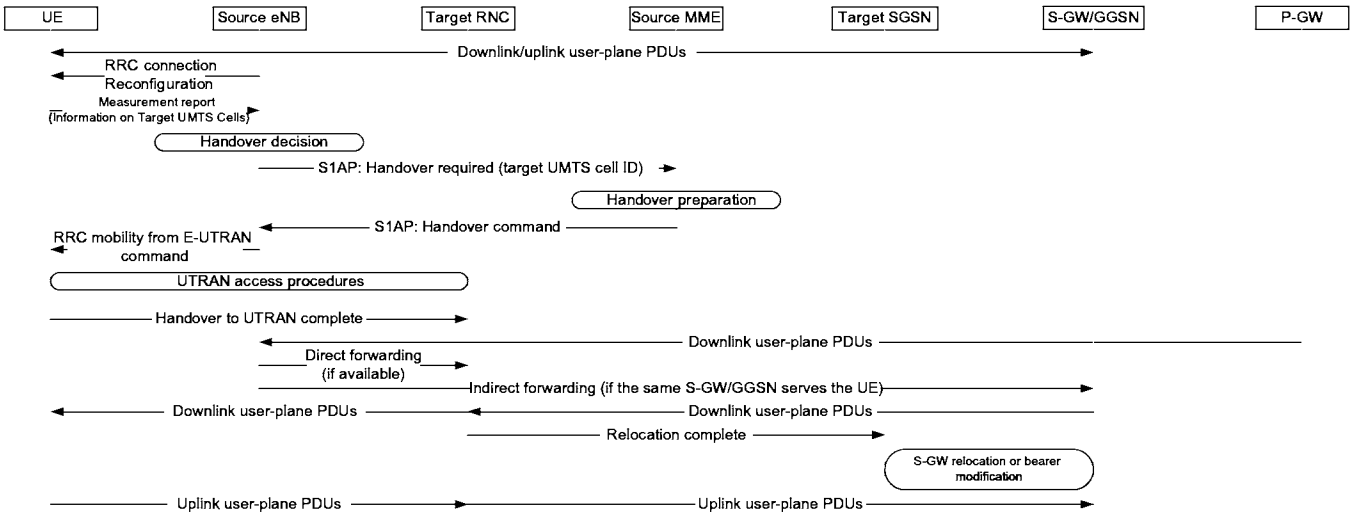


FIGURE 4.29

E-UTRA to UMTS handover procedure [29].

4.8 Overview of IP packet processing procedures

The IP packets are mapped to one or more radio bearers that are provided by E-UTRAN according to the QoS requirements of the corresponding services. An overview of IP packet processing stages in the E-UTRA downlink is illustrated in Figure 4.30. It must be noted that depending on the content

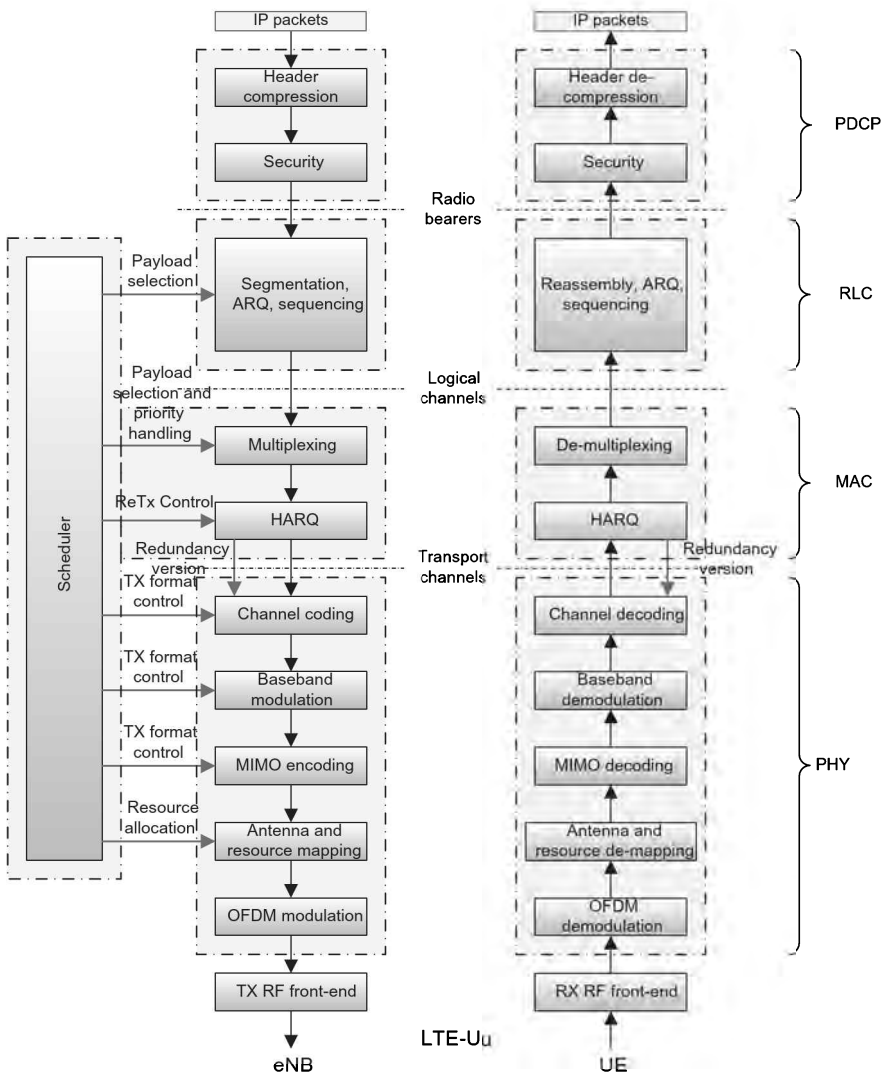


FIGURE 4.30

Overall downlink IP packet processing procedures [22].

of the packets, some of the functions shown in this figure may not be applicable. For example, MAC scheduling and HARQ soft combining functions are not relevant to broadcast of the system information or other control-plane packets. The uplink data packet processing is similar to the downlink with the exception of some differences related to transmission format selection (modulation and coding scheme, transmission modes, multi-antenna processing, etc.) due to dissimilarities in downlink/uplink multiple access schemes and other physical-layer functions.

As shown in Figure 4.30, the IP packets arriving at the PDCP sublayer are processed through the header-compression module in order to make efficient use of radio resources by reducing the number of redundant bits transmitted over the radio interface. The header-compression mechanism is based on the robust header compression (ROHC) scheme [18]. The PDCP sublayer is also responsible for ciphering and, for the control-plane packets, integrity protection of the transmitted data, as well as in-sequence delivery and duplicate removal for handover. A PDCP header is added to PDCP SDUs, carrying information required for deciphering of data in the terminal. At the receiver side, the PDCP performs the corresponding deciphering and decompression functions. There is one PDCP entity per radio bearer configured for a terminal [8].

The PDCP PDUs then go through the RLC sublayer which is responsible for segmentation or concatenation, retransmission, duplicate detection, and in-sequence delivery of the packets to the higher layers. The RLC provides services to the PDCP in the form of radio bearers. There is one RLC entity per radio bearer configured for a terminal. The addition of an RLC header facilitates in-sequence delivery (per logical channel) of the PDCP SDUs in the terminal and identification of the RLC PDUs in the case of retransmissions. The RLC PDUs are forwarded to the MAC sublayer, which multiplexes a number of RLC PDUs and attaches a MAC header to form a transport block. The transport-block size depends on the instantaneous rates selected by the link adaptation mechanism, which affects both MAC and RLC processing [7].

The MAC sublayer performs multiplexing of logical channels, HARQ retransmissions, and uplink/downlink scheduling services. Note that the scheduling function is only present in the eNB for both uplink and downlink packet scheduling services. The HARQ protocol is present in both transmitting and receiving ends of the MAC sublayer. The MAC sublayer provides services to the RLC sublayer in the form of logical channels [6].

The physical layer performs encoding/decoding, modulation/demodulation, scrambling/de-scrambling, interleaving/de-interleaving, multi-antenna processing, resource mapping/de-mapping, and OFDM/SC-FDMA modulation/de-modulation functions. The physical layer provides services to the MAC sublayer in the form of transport channels. The physical layer attaches a CRC to the transport block for error detection purposes, performs coding and modulation, and transmits the resulting signal using multiple transmit antennas [22].

Samsung Ex. 1005, Page 254 of 1147

References

3GPP Specifications[12]

[1] 3GPP TS 23.003, Numbering, Addressing and Identification (Release 11), December 2012.

[2] 3GPP TS 29.274, Evolved General Packet Radio Service (GPRS) Tunneling Control Protocol for Control Plane (GTPv2-C) Stage 3 (Release 11), December 2012.

[3] 3GPP TS 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 11), December 2012.

[4] 3GPP TS 24.301, Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3 (Release 11), December 2012.

[5] 3GPP TS 36.304, Evolved Universal Terrestrial Radio Access (E-UTRA); UE Procedures in Idle Mode (Release 11), December 2012.

[6] 3GPP TS 36.321, Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification (Release 11), December 2012.

[7] 3GPP TS 36.322, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification (Release 11), December 2012.

[8] 3GPP TS 36.323, Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) specification (Release 11), December 2012.

[9] 3GPP TS 36.331, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 11), December 2012.

[10] 3GPP TS 23.401, General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (Release 11), March 2012.

[11] 3GPP TS 23.402, Architecture Enhancements for non-3GPP Accesses (Release 11), December 2012.

[12] 3GPP TS 36.413, Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S1 Application Protocol (S1AP) (Release 11), December 2012.

[13] 3GPP TR 36.839, Evolved Universal Terrestrial Radio Access (E-UTRA); Mobility Enhancements in Heterogeneous Networks (Release 11), December 2012.

3GPP2 Specifications[13]

[14] 3GPP2 C.S0024-A v3.0, cdma2000 High Rate Packet Data Air Interface Specification, September 2006.

[15] 3GPP2 C.S0005-A v6.0, Upper Layer (Layer 3) Signaling Standard for cdma2000 Spread Spectrum Systems, February 2002.

[16] 3GPP2 C.S0004-A v6.0, Signaling Link Access Control (LAC) Standard for cdma2000 Spread Spectrum Systems, February 2002.

[12] 3GPP specifications can be found at the following URL: http://www.3gpp.org/ftp/Specs/archive/.

[13] 3GPP2 specifications can be found at the following URL: http://www.3gpp2.org/Public_html/specs/index.cfm.

IETF Specifications[14]

[17] IETF RFC 2460, S. Deering, Internet Protocol, Version 6 (IPv6) Specification, December 1998.

[18] IETF RFC 3095, C. Bormann, RObust Header Compression (ROHC): Framework and four profiles: RTP, UDP, ESP, and uncompressed, July 2001.

[19] IETF RFC 6733, V. Fajardo, Diameter Base Protocol, October 2012.

Books

[20] S. Sesia, I. Toufik, M. Baker, *LTE—The UMTS Long Term Evolution: From Theory to Practice*, second ed., John Wiley & Sons, 2011.

[21] C. Johnson, Long Term Evolution in Bullets, second ed., Create Space Independent Publishing Platform, 2012.

[22] E. Dahlman, S. Parkvall, J. Skold, 4G, LTE/LTE-Advanced for Mobile Broadband, Academic Press, 2011.

Articles

[23] LTE and Beyond: LTE, 4G, MME, PGW, SGW, interfaces and beyond. <http://www.lteandbeyond.com/>.

[24] H. Tao, Z. Zhijiang, L. Yunjie, *QoSmechanism in EPS*, ZTE Commun. (2009).

[25] V. Srinivasa Rao, R. Gajula, Protocol signaling procedures in LTE, Radisys Corp. (2011).

[26] V. Srinivasa Rao, R. Gajula, Interoperable UE handovers in LTE, Radisys Corp. (2011).

[27] GPRS tunneling protocol. <http://en.wikipedia.org/wiki/GPRS_Tunnelling_Protocol>.

[28] The LTE Network Architecture: A comprehensive tutorial, Alcatel-Lucent, 2009.

[29] J. Laganier, et al., Mobility management for All-IP core network, NTT DoCoMo Tech. J. 11 (3) (2009).

[30] K. Suzuki, et al., Core network (EPC) for LTE, NTT DoCoMo Tech. J. 13 (3) (2011).

[31] LTE Identifiers, NMC Consulting Group. <http://www.nmcgroups.com>, 2011.

[32] LTE system description. <http://www.sharetechnote.com/>.

[33] A. Roessler, Cell Search and Cell Selection in UMTS LTE, Rohde & Schwarz, 2009.

[34] Agilent Technologies, 3GPP Long Term Evolution: System Overview, Product Development, and Test Challenges, September 2009.

[35] 3G and 4G wireless blog <http://3g4g.blogspot.com/>.

[36] W. Brouwer, QoS in LTE, Alcatel Lucent, December 2010.

[37] S. Amberkar, Telecom Tutorials, <http://samiramberkar-tutorials.blogspot.com/>.

[38] 4G Americas White Paper, developing and integrating a high performance hetnet. <http://www.4gamericas.org/>, October 2012.

[14]IETF specifications can be found at the following URL: http://www.ietf.org/rfc.html.

This page intentionally left blank

CHAPTER

Radio Resource Control Functions

5

CHAPTER CONTENTS

5.1 RRC states and state transitions 230
5.1.1 RRC_CONNECTED state 232
5.1.2 RRC_IDLE state 234
5.2 RRC connection management 234
5.2.1 Signaling and data radio bearers 240
5.2.2 Quality of service control 244
5.3 RRC mobility management 246
5.3.1 Mobility in RRC_CONNECTED state 251
5.3.2 Mobility in RRC_IDLE state 252
5.3.3 Intra-RAT mobility and handover 253
5.3.4 Inter-RAT mobility and handover 255
5.4 Measurements and measurement reporting 257
5.5 RRC procedures 262
5.5.1 System information 262
5.5.2 Paging 266
5.5.3 PLMN and cell selection 275
5.5.4 Cell reselection 278
5.5.5 Radio link failure 281
5.6 RRC timers and constants 282
5.7 RRC message structure 282
References 286

The radio resource control (RRC) sublayer in E-UTRAN protocol stack performs control-plane functions such as broadcast of system information related to access stratum (AS) and non-access stratum (NAS); paging; establishment, maintenance and release of an RRC connection between the UE and E-UTRAN; signaling radio bearer (SRB) management; security handling; mobility management including UE measurement reporting and configuration; connected mode

LTE-Advanced.

© 2014 Elsevier Inc. All rights reserved.

Samsung Ex. 1005, Page 258 of 1147

handover; idle-mode mobility control; multimedia broadcast multicast service (MBMS) notification services and radio bearer management for MBMS; quality of service (QoS) management and NAS direct message transfer from NAS to the UE and vice versa [1]. Figure 5.1 illustrates the RRC sublayer and its interactions with other protocol layers in the E-UTRA protocol stack in terms of configuration of lower-layer functions and transfer of upper-layer control/signaling data via lower layers. The RRC sublayer provides the upper layers with the following services: (1) broadcast of common control information; (2) notification of idle-mode UEs and (3) transfer of dedicated control information. At the same time, the RRC sublayer receives the following services from lower layers: (1) integrity protection and ciphering from packet data convergence protocol (PDCP) sublayer; and (2) reliable and in-sequence transfer of information,

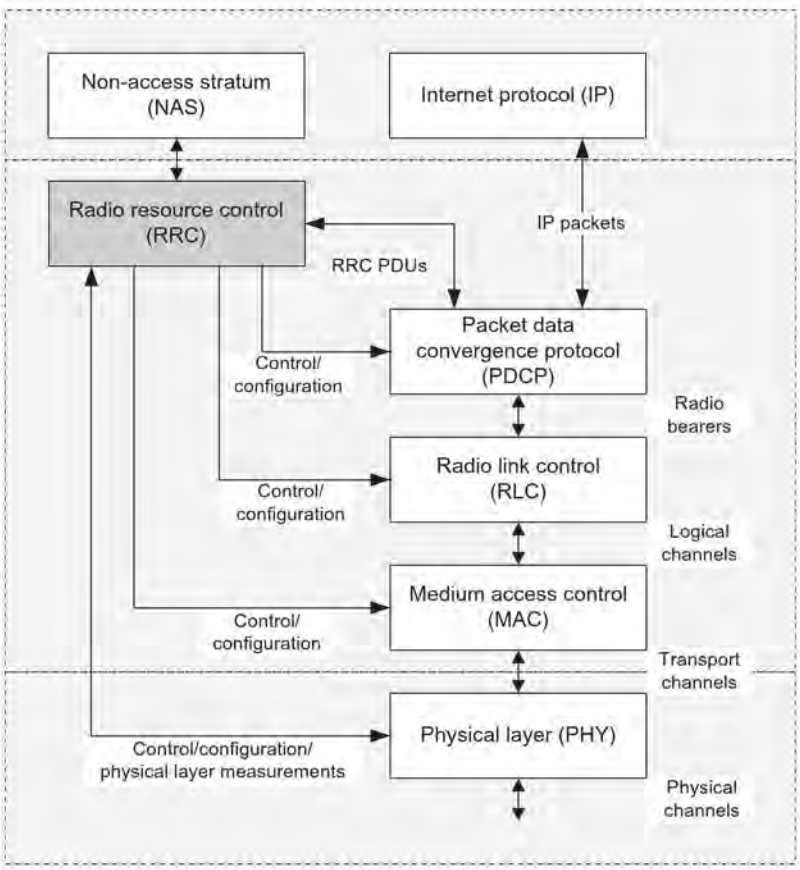


FIGURE 5.1

The RRC sublayer in the LTE protocol stack.

without introducing duplicates and support for segmentation and concatenation from the radio link control (RLC) sublayer.

More specifically, the RRC sublayer performs the following functions [1]:

- Broadcast of system information including NAS common information and the information pertinent to UEs in RRC_IDLE state such as cell selection/reselection parameters, neighbor cell information, and information related to UEs in RRC_CONNECTED state such as common channel configuration information.
- RRC connection control that includes the following tasks:
  - Paging of the UEs in the idle mode.
  - Establishment, modification, or release of the RRC connection which includes assignment and/or modification of the UE identity; establishment, modification, or release of SRBs; and access class barring.
  - Initial security activation which includes initial configuration of AS integrity protection and ciphering [3].
  - Maintaining RRC connection while the UEs are moving which includes intra-frequency/inter-frequency handover-associated security handling, i.e., key/algorithm change, specification of RRC context information transferred between network nodes.
  - Establishment, modification, or release of data radio bearers (DRBs).
  - Radio configuration and control in the lower layers including automatic repeat request (ARQ), hybrid automatic repeat request (HARQ), and discontinuous reception (DRX) configuration.
  - Cell management including change of primary cell; addition, modification, or release of secondary cells; and addition, modification, or release of secondary timing advanced groups when carrier aggregation is supported.
  - QoS control including assignment and modification of downlink/uplink semi-persistent scheduling configuration; assignment and modification of parameters for uplink rate control in the UE, i.e., allocation of a priority and a prioritized bit rate for each radio bearer.
  - Recovery from radio link failure (RLF).
- Inter-radio access technology (inter-RAT) mobility including security activation and transfer of RRC context information.
- Measurement configuration and reporting, which includes establishment, modification, or release of intra-frequency/inter-frequency/inter-RAT measurements; setup and release of measurement gaps; and measurement reporting.
- Support of self-configuration and self-optimization as well as support of measurement logging and reporting for network performance optimization.

In order to simplify the operation and to reduce state transition latency, LTE simplified the legacy systems and defined two operational states for the UE: RRC_CONNECTED and RRC_IDLE. A UE is in the RRC_CONNECTED state when an RRC connection or control-plane has been established; otherwise, the UE is said to be in the RRC_IDLE state. The UE states only correspond to the

steady-state behavior and the transient procedures for cell search and selection, as well as network attach/detach, are not designated as states.

In this chapter, we will take a top-down systematic approach to analyzing and understanding the services and functions of the RRC sublayer in LTE/LTE-Advanced and will provide practical and implementation perspectives on LTE RRC sublayer protocols.

5.1 RRC states and state transitions

In LTE, a user terminal at any given time during normal operation can be in one of the two (steady-state) RRC states as shown in Figure 5.2. The emphasis on the steady-state descriptor is due to the fact that the state diagram shown in Figure 5.2 does not include the transient states of the mobile terminal. In fact, when a UE is powered on or upon RLF, it goes to a transient state with no attachment to a public land mobile network (PLMN) or a cell, and thus has no connection to the network. The UE starts scanning and selecting a PLMN and then proceeds to camp on a cell (cell selection/reselection) through downlink synchronization and system information acquisition. In this case, the UE has virtually moved to the RRC_IDLE state since it still has no RRC connection and thus performs uplink synchronization and RRC connection setup procedures. The RRC states can further be characterized as follows [1]:

- RRC_IDLE
  - UE-specific DRX is configured by NAS.
  - UE-controlled mobility is performed.
  - UE monitors a paging channel to detect incoming calls.
  - System information change is tracked and earthquake and tsunami warning system (ETWS) notification is monitored.
  - UE performs neighbor cell measurements and cell selection or reselection and acquires the system information.

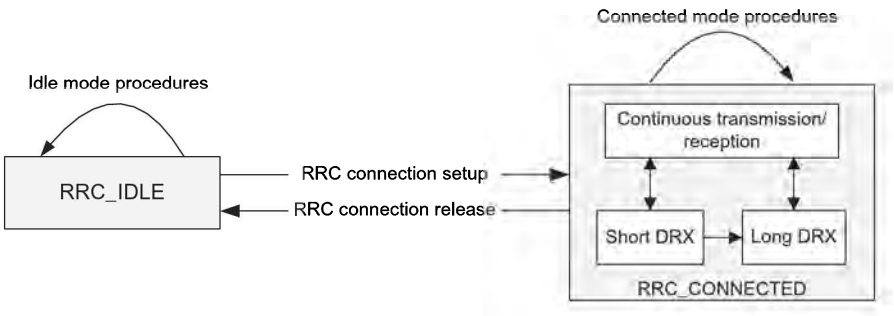


FIGURE 5.2

UE (steady-state) RRC states in LTE/LTE-Advanced.

- RRC_CONNECTED
  - Bi-direction transfer of unicast data in the uplink and downlink.
  - At lower layers, the UE may be configured with a UE-specific DRX.
  - Network-controlled mobility, i.e., handover and cell change order (CCO) with optional network-assisted cell change to GERAN.
  - Neighbor cell measurements.

In RRC_IDLE, the UE monitors a paging channel and/or system information block type 1 (SIB1) content to detect system information change, and for ETWS capable UEs, ETWS notifications. The UE further monitors control channels associated with the shared data channel to determine whether data is scheduled for it.

Prior to transition to RRC steady-states, the UE uses the information stored on its universal subscriber identity module (USIM) and the mobile equipment (ME) for system selection and camping. Upon initial power-up, the UE uses the static information on the USIM and ME to initiate system selection. In a practical scenario, the international mobile subscriber identity (IMSI) on the USIM may be used to determine the home PLMN code, and may serve as the basis for PLMN selection. The USIM may have optionally an equivalent home PLMN list and priorities among them to determine the PLMN selection. The USIM may also provide the allowed radio access technology (RAT) for each PLMN where there can be multiple RATs for a given PLMN. The UE considers the supported RAT and frequency band to trigger system selection. The system selection procedures are divided between NAS and AS. Over the air, signaling support provided by the AS assists NAS in camping on a PLMN. The selected PLMN is known as the registered PLMN and is stored on the USIM for future system selections.

If a PLMN is selected, but the UE cannot register with it because the registration is rejected, the PLMN is added to the prohibited PLMN list, which is retained even when the UE is switched off or the USIM is removed. A prohibited PLMN, which is typically a visited PLMN, can usually be selected only with user manual intervention. The ME uses the registered PLMN or a list of equivalent PLMNs (obtained during registration, all considered equal priority PLMNs) for PLMN selection (not for initial power-up), cell selection/reselection, and handover. If registration cannot be completed on any PLMN, the UE indicates a *No Service* to the user and waits until a new PLMN is detected, or tracking areas (TAs) of an allowed PLMN that are not in the prohibited TA list are found. The UE selects the registered PLMN or equivalent PLMN to initiate system selection during recovery from RLF or at power-up.

The selection of a PLMN can be manual or automatic. In both cases, the UE scans all possible RF channels that it supports based on its capability [5]. During the search, all detected PLMNs with reference signal received power (RSRP) ≥ -110 dBm are reported to the NAS as high-quality PLMNs without an associated RSRP value. The PLMNs with RSRP < -110 dBm are reported along with an associated RSRP value [10]. The order in which the UE searches for PLMNs based on either intra- or inter-band is not defined but can be

optimized based on some information that is available to the UE including the last frequency on which it successfully camped. In both automatic and manual modes, the search for PLMNs can be discontinued by the NAS at any time. When the UE is in automatic mode and multiple PLMNs are detected, the order in which the selection is made is controlled by a set of rules such as last-registered PLMN (or equivalent) which is initially attempted, then home (or equivalent) PLMN is tried, followed by a user-defined and/or operator-defined list of PLMNs in priority order, followed by high-quality PLMNs in random order, followed by other PLMNs in the order of decreasing RSRP. If a UE is camped on a PLMN that is not the highest priority, it will periodically search for higher priority PLMNs and report the results to the NAS. The time interval between searches is stored on the USIM and set by the service provider to be between 6 minutes and 8 hours, in increments of 6 minutes.

5.1.1 RRC_CONNECTED state

The RRC_CONNECTED state is a state where the UE transfers unicast control/data in the uplink or downlink. The UE may be configured with a customized DRX or discontinuous transmission (DTX) pattern, or both. The UE monitors control channels associated with the shared data channel to determine if data is scheduled for it, provides channel quality and feedback information, performs neighbor cell measurements and measurement reporting, and acquires the system information. One or multiple IP addresses are assigned to the UE as well as a temporary identifier called cell radio network temporary identifier (C-RNTI) that is used for signaling purposes between the UE and the network. In brief, the RRC_CONNECTED state is characterized as follows:

- UE is known in EPC and E-UTRAN/eNB, i.e., UE context is stored in eNB.
- UE location is known at cell level.
- Unicast data transfer with UE is possible
- Short/long DRX is supported for power saving.
- Mobility is controlled by the network.

In RRC_CONNECTED, the network controls UE mobility, i.e., the network decides when the UE must move to another cell, which may be on another frequency or RAT. For network-controlled mobility in the RRC_CONNECTED state, handover is the only procedure that is defined. The network triggers the handover procedure based on radio conditions or network load. The network may configure the UE to perform measurement reporting including the configuration of measurement gaps. The network may also initiate handover without receiving measurement reports from the UE. Before sending the handover message to the UE, the source eNB prepares one or more target cells. The target eNB generates the message used to perform the handover, including the AS configuration to be used in the target cell. The source eNB transparently forwards the handover message/information received from the target to the UE. The source eNB may initiate

data forwarding for some of the user DRBs. After receiving the handover message, the UE attempts to access the target cell at the first available random access channel (RACH) opportunity according to the random access resource selection procedure for asynchronous handover. When allocating a dedicated preamble for the random access in the target cell, the E-UTRA ensures it is available from the first RACH opportunity that the UE may utilize. Upon successful completion of the handover, the UE sends a message used to confirm the handover. Following successful completion of the handover, the PDCP service data units (SDUs) may be retransmitted in the target cell. This only applies for user DRBs using radio link control acknowledged mode (RLCAM) mode. Furthermore, the sequence number (SN) and the hyper-frame number (HFN)[1] are reset, except for the user DRBs using RLCAM mode for which both SN and HFN continue. For a limited period of time, the source eNB maintains the UE context to allow the UE to return in case of handover failure. If handover failure is detected, the UE will attempt to resume the RRC connection with either the source eNB or another cell using the RRC connection reestablishment procedure. The connection reestablishment would succeed, if the accessed cell is prepared.

The paging procedure is used to transmit paging information to a UE in RRC_IDLE and/or to inform mobile terminals in the RRC_IDLE and RRC_CONNECTED states about a system information change and/or to inform about an ETWS primary and/or secondary notification. The paging information is provided to upper layers, which in response may establish an RRC connection to receive an incoming call. The E-UTRAN initiates the paging procedure by transmitting the paging message at the UE's paging opportunities. The E-UTRAN may address multiple mobile terminals within a paging message by including one paging record for each UE. The E-UTRAN may also indicate a change of system information and/or provide an ETWS notification.

The RRC connection establishment involves SRB1 establishment. The procedure is also used to transfer the initial NAS dedicated information/message from the UE to the E-UTRAN. The UE initiates the RRC connection establishment procedure when the upper layers request establishment of an RRC connection while the UE is in the RRC_IDLE state.

The RRC connection reestablishment is used to reestablish the RRC connection including resumption of SRB1 operation and the reactivation of security. A UE in RRC_CONNECTED, for which security has been activated, may initiate the procedure in order to continue the RRC connection. The connection reestablishment succeeds, if the serving/target is prepared, i.e., it has a valid UE context. If the E-UTRAN accepts the connection reestablishment, SRB1 operation resumes while the operation of other radio bearers is suspended. If the AS security has not

[1]An HFN is an overflow counter mechanism that is used in the eNB and UE in order to limit the actual number of sequence number bits that are needed to be sent over the radio air-interface. The HFN needs to be synchronized between the UE and eNB.

Table 5.1 Summary of the Differences Between MAC and RRC Control [1]

	MAC Control		RRC Control
Control Entity	**MAC**		**RRC**
Signaling type	PDCCH	MAC control PDU	RRC message
Signaling reliability	$\sim 10^{-2}$ (no retransmission)	$\sim 10^{-3}$ (with HARQ)	$\sim 10^{-6}$ (with ARQ)
Control latency	Very short	Short	Long
Extensibility	None	Limited	High
Security	No integrity protection; no ciphering	No integrity protection; no ciphering	Integrity-protected ciphering

been established, the UE cannot initiate the procedure and directly transitions to the RRC_IDLE state.

The main difference between medium access control (MAC) and RRC control lies in the signaling reliability. The signaling corresponding to state transitions and radio bearer configurations should be performed by the RRC sublayer due to signaling reliability. The different characteristics of MAC and RRC control are summarized in Table 5.1.

5.1.2 RRC_IDLE state

The RRC_IDLE is a state where a UE-specific DRX may be configured by upper layers. In the idle mode, the UE saves power and does not inform the network of each cell change. The network knows the location of the UE to the granularity of a few cells, known as the tracking area. The UE monitors a paging channel to detect incoming traffic, performs neighbor cell measurements and cell selection/reselection, and acquires system information. In brief, the RRC_IDLE state is characterized as follows [1,4]:

- UE is known in EPC and has an assigned IP address.
- UE is not known in E-UTRAN/eNB.
- User terminal's location is known at tracking-area level.
- Unicast data transfer with UE is not possible.
- UE can be reached through paging in TAs controlled by EPC.
- UE-based cell selection and TA update are provided to EPC.

5.2 RRC connection management

One of the main functions of the RRC sublayer is to establish, maintain, modify, and release connections with E-UTRAN for transport of user data and control

messages. There are two NAS states that reflect the state of a UE with respect to connection establishment: (1) evolved packet system (EPS) mobility management (EMM) state (EMM-DEREGISTERED or EMM-REGISTERED) which reflects whether the UE is registered in the mobility management entity (MME); and (2) EPS connection management (ECM) state (ECM-IDLE or ECM-CONNECTED) which signifies the connectivity of the UE with the core network.

The transition from ECM-IDLE to ECM-CONNECTED not only involves establishment of the RRC connection, but also includes establishment of S1 connection. The RRC connection establishment is initiated by the NAS and is completed prior to S1 connection establishment, which means that connectivity in RRC_CONNECTED is initially limited to the exchange of control information between UE and E-UTRAN. The UEs are typically moved to ECM-CONNECTED when becoming active. It should be noted that in LTE, the transition from ECM-IDLE to ECM-CONNECTED is performed within 100ms. The UEs engaged in intermittent data transfer do not need to stay in ECM-CONNECTED, if the ongoing services can tolerate such transfer delays. The LTE design objective was to support similar battery power consumption levels for UEs in RRC_CONNECTED as UEs in RRC_IDLE. The RRC connection release is initiated by the eNB following release of the S1 connection between the eNB and the core network [19].

The RRC connection establishment starts with the establishment of SRB1 and the transfer of the initial NAS message. This NAS message triggers the setup of an S1 connection, which normally initiates a subsequent step during which E-UTRAN activates AS security and further establishes SRB2, along with one or more DRBs associated with the default and optionally dedicated EPS bearers. Figure 5.3 illustrates the RRC connection establishment and the transport paths of user application data, NAS signaling, and RRC control messages. Note that NAS messages (ECM/EMM) are encapsulated in RRC messages (Figure 5.4), thus the RRC connection establishment needs to happen prior to NAS signaling transfer. Furthermore, it must be noted that SRBs are established before DRBs. The integrity protection algorithm in LTE/LTE-Advanced is common for signaling radio bearers SRB1 and SRB2. The ciphering algorithm is common for all radio bearers (i.e., SRB1, SRB2, and DRBs). However, neither integrity protection nor ciphering applies to SRB0 [1].

More specifically, the RRC connection establishment involves the establishment of SRB1. E-UTRAN completes the RRC connection establishment prior to establishing an S1 connection, i.e., prior to receiving the UE context from EPC. As a result, AS security is not activated during the initial phase of the RRC connection.

Upon receiving the UE context from the EPC, the E-UTRAN activates security that includes ciphering and integrity protection using the initial security activation procedure. The RRC messages to activate security are integrity-protected, while ciphering is started only after completion of the procedure. That is, the response to the message used to activate security is not ciphered, while the

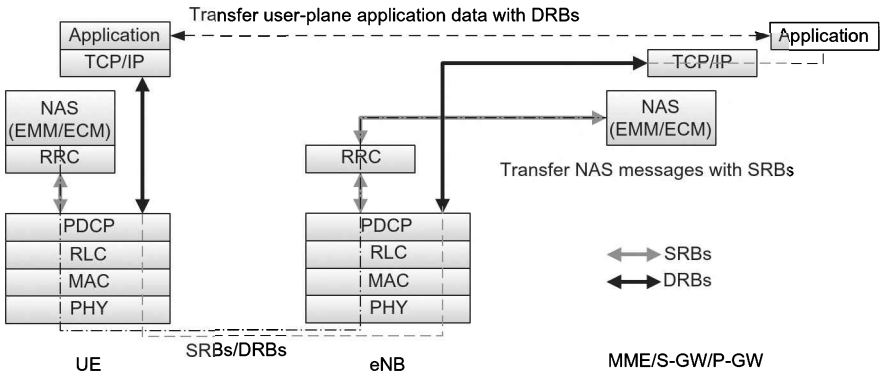


FIGURE 5.3

Illustration of different radio bearers in LTE RRC_CONNECTED state.

subsequent messages used to establish SRB2 and DRBs are both integrity-protected and ciphered. Following initiation of the security activation procedure, E-UTRAN proceeds to establish SRB2 and DRBs. The E-UTRAN will apply ciphering and integrity protection for the RRC connection reconfiguration messages used to establish SRB2 and DRBs. The E-UTRAN should release the RRC connection, if the initial security activation and/or the radio bearer establishment fail. For SRB2 and DRBs, security is always activated from the beginning, i.e., the E-UTRAN does not establish these bearers prior to activating security.

Once the security activation procedure is performed, the E-UTRAN may configure a carrier-aggregation-enabled UE with one or more secondary cells in addition to the primary cell that was initially configured during RRC connection establishment. The primary cell is used to provide the security inputs and upper-layer system information. The secondary cells are used to provide additional downlink and optionally uplink radio resources for user data transmission [1].

In order to better understand the RRC connection configuration/reconfiguration/release process, we need to briefly review the EPS mobility and connection management states. The EMM states describe the mobility management states that result from the mobility management procedures, e.g., attach and TA update procedures. There are two EMM states defined in LTE: EMM-DEREGISTERED and EMM-REGISTERED. The ECM states describe the signaling connectivity between the UE and the EPC. LTE defines two ECM states: ECM-IDLE and ECM-CONNECTED.

In general, the ECM and EMM states are independent of each other. Transition from EMM-REGISTERED to EMM-DEREGISTERED can occur regardless of the ECM state, e.g., by explicit detach signaling in ECM-CONNECTED or by

Samsung Ex. 1005, Page 267 of 1147

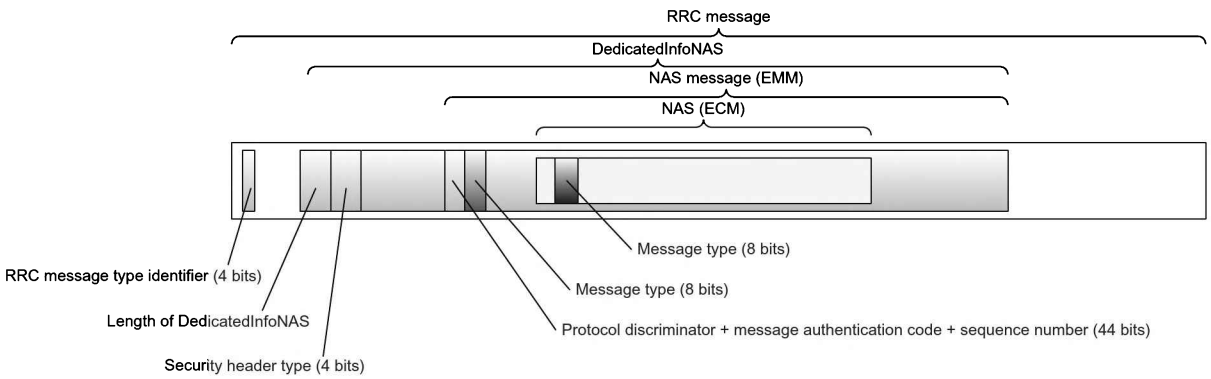


FIGURE 5.4

Encapsulation of NAS signaling/messages in RRC messages [27].

implicit detach locally in the MME during ECM-IDLE. However, in order to transition from EMM-DEREGISTERED to EMM-REGISTERED, the UE has to be in the ECM-CONNECTED state. In the EMM DEREGISTERED state, the EMM context in MME holds no valid location or routing information for the UE. The UE is not reachable by an MME, as the UE location is not known. In the EMM-DEREGISTERED state, some UE context can still be stored in the UE and MME, e.g., to avoid running a key agreement (AKA) procedure during every attach procedure.

A UE is in ECM-IDLE state when no NAS signaling connection between the UE and the network exists. In the ECM-IDLE state, a UE performs cell selection/reselection and PLMN selection. There is no UE context in E-UTRAN for the UE in the ECM-IDLE state. There is no S1-MME and no S1-U connection for the UE in the ECM-IDLE state. The UE location is known in the MME with an accuracy of a tracking area. The UE performs the TA update procedure when the tracking area identity (TAI) in the EMM system information is not in the list of TAs that the UE registered with the network, or when the UE performs a handover to an E-UTRAN cell and the UE's temporary identity indicates P-TMSI. For a UE in the ECM-CONNECTED state, there is a signaling connection between the UE and the MME. The signaling connection consists of two parts: an RRC connection and an S1-MME connection [7].

The RRC connection establishment process has been illustrated in Figure 5.5. The UE attempts to synchronize with the downlink and acquire the system information. The NAS layer in the UE (see Figure 5.3) triggers connection establishment, which may be in response to a paging message. If the access is not barred, the lower layers in the UE perform a contention-based random access procedure. If the contention-based random access procedure is successful, the UE sends the *RRC Connection Request* message. This message includes an initial identity, which is typically the UE's system architecture evolution temporary mobile subscriber identity (S-TMSI) or a random number. If E-UTRAN accepts the connection request, it returns the *RRC Connection Setup* message that includes the initial radio resource configuration including SRB1. Instead of signaling each individual parameter, the E-UTRAN may order the UE to apply a default configuration, i.e., a configuration for which the parameter values are specified in the RRC specification [1]. The UE responds with the *RRC Connection Setup Complete* message with an encapsulated NAS message, an identifier of the selected PLMN, and an identifier of the registered MME. Based on the last two parameters, the eNB decides with which MME it should establish the S1 connection.

The E-UTRAN sends the *Security Mode Command* message to activate integrity protection and ciphering. This message, which is integrity-protected but not ciphered, indicates which algorithms must be used. The UE verifies the integrity protection of the *Security Mode Control* message and configures lower layers to apply integrity protection and ciphering to all subsequent messages with the exception that ciphering is not applied to the response message, i.e., the *Security Mode*

Samsung Ex. 1005, Page 269 of 1147

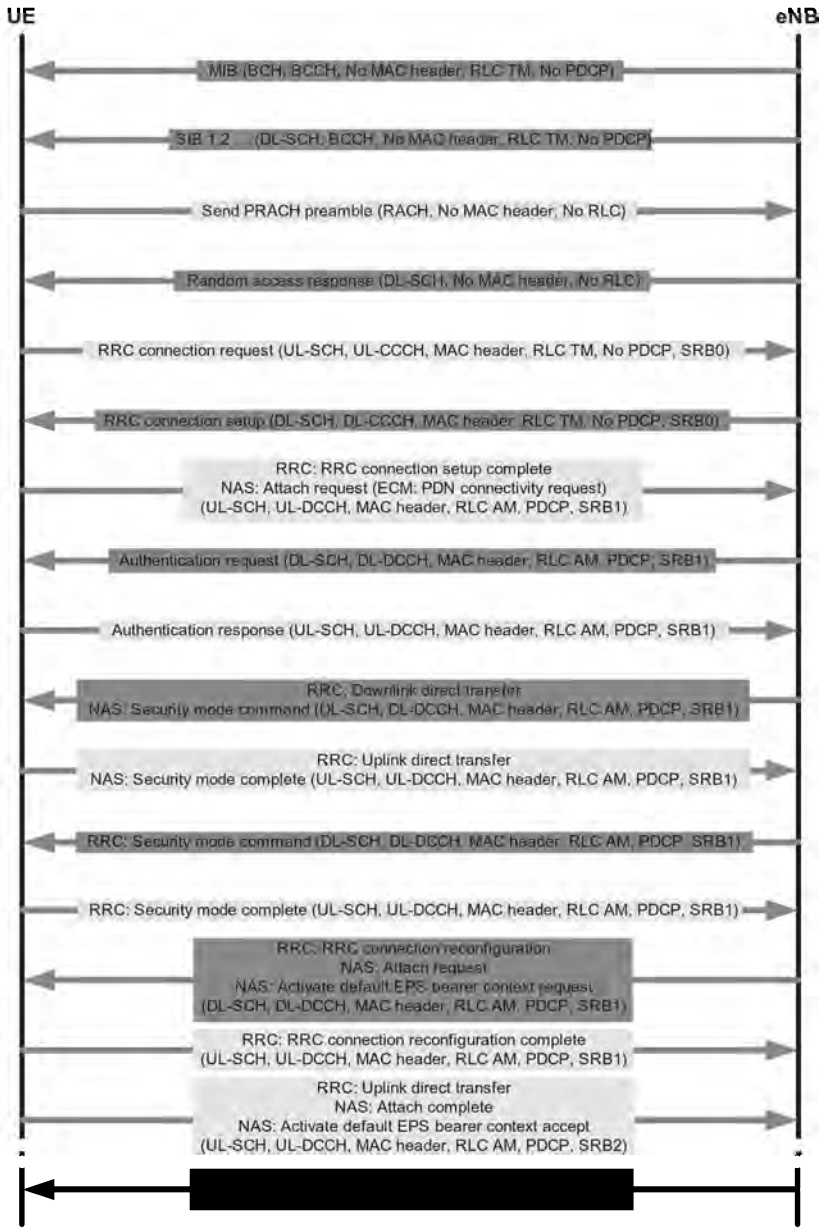


FIGURE 5.5

Sequence of signaling for RRC connection setup, reconfiguration, and release [27].

Complete or *Security Mode Failure* messages are not ciphered. The E-UTRAN sends the message including a radio resource configuration used to establish SRB2 and one or more DRBs. This message may also include other information such as a piggybacked (encapsulated) NAS message or a measurement configuration. The E-UTRAN may send the *RRC Connection Reconfiguration* message prior to receiving the *Security Mode Complete* message. In this case, the E-UTRAN should release the connection when one or both procedures fail. The UE ultimately returns the *RRC Connection Reconfiguration Complete* message [19,27].

5.2.1 Signaling and data radio bearers

A signaling radio bearer is defined as a radio bearer that is only used for transmission of RRC and NAS messages. More specifically, the following three SRB types are defined in E-UTRA [1]:

- SRB0 is used for transport of RRC messages associated with the common control logical channel.
- SRB1 is used for transport of RRC messages including piggybacked NAS messages as well as NAS messages prior to the establishment of SRB2 where all are associated with a dedicated control logical channel.
- SRB2 is used for transport of NAS messages using a dedicated control logical channel, has a lower priority than SRB1, and is always configured by E-UTRAN after security activation.

The SRB types and their usage are summarized in Table 5.2. The UE in the RRC_IDLE state does not have any allocated SRB or DRB. The UE requires (dedicated) SRBs and DRBs to exchange data/messages with the network. In particular, for transfer of NAS messages, SRBs are established. They carry registration and signaling for user-plane setup. The user-plane application data is transported on DRBs. An E-UTRAN radio access bearer (E-RAB) uniquely identifies the concatenation of an S1 bearer and the corresponding DRB. When an E-RAB exists, there is a one-to-one mapping between this E-RAB and an EPS

Table 5.2 E-UTRA SRBs and Their Usage [1]

SRB	Message Type	Associated Logical Channel
SRB0	RRC messages (*it is used when the RRC connection is established*)	CCCH
SRB1	RRC messages, RRC messages with piggybacked NAS messages, NAS messages prior to SRB2 setup	DCCH
SRB2	NAS messages, RRC messages with piggybacked NAS messages (*lower priority than SRB1*)	DCCH

bearer of the NAS [4]. To establish, modify, or release DRBs, the E-UTRAN applies the RRC connection reconfiguration procedure.

DRBs are established by the eNB in the RRC_CONNECTED state based on the relevant information from the MME. This information includes the desired QoS for the radio bearer, which enables the eNB to appropriately configure the PDCP, RLC, and MAC parameters. In the RRC_CONNECTED state, the DRBs can be modified and released based on the signaling information received from the MME. Some restrictions may apply with regard to the modification of existing DRBs. All DRBs are implicitly released when the RRC connection is released. When establishing a DRB, the E-UTRAN decides how to transfer the packets of an EPS bearer across the radio interface. An EPS bearer is uniquely mapped to a DRB, a DRB is then mapped to a dedicated traffic logical channel, all logical channels are mapped to a downlink or uplink shared transport channel, which are mapped to the corresponding physical downlink or uplink shared channel. A UE in the RRC_CONNECTED state, for which security has been activated, may initiate the procedure in order to continue the RRC connection. The connection reestablishment succeeds only if the serving cell is prepared, i.e., it has a valid UE context. In the case where the E-UTRAN accepts the connection reestablishment, SRB1 operation resumes while the operation of other radio bearers remains suspended. If the AS security has not been activated, the UE does not initiate the procedure, but rather directly transitions to the RRC_IDLE state.

Before concluding this section, it would be useful if we summarize the EPS bearer architecture and understand the termination points of each bearer in the LTE system. The EPS bearer service architecture is illustrated in Figure 5.6, in which the following bearers are defined [1,4]:

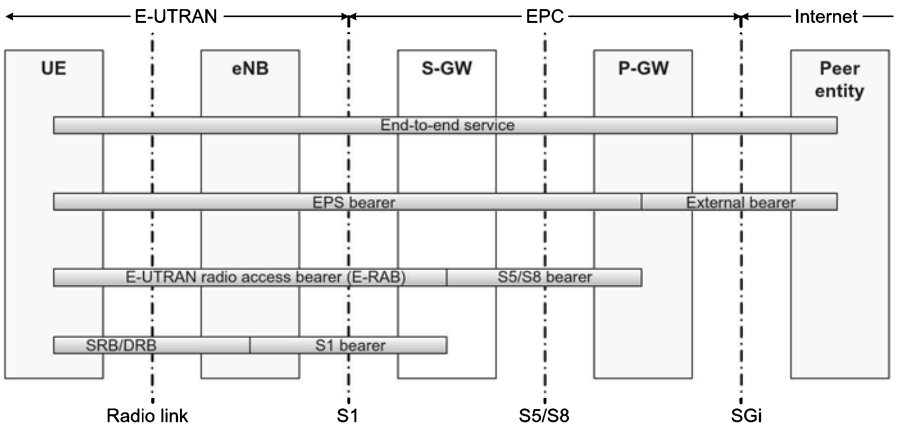


FIGURE 5.6

EPS bearer service architecture [4].

- An uplink traffic flow template (TFT) binds a service data flow to an EPS bearer in the uplink direction, where multiple service data flows can be multiplexed into the same EPS bearer by including multiple uplink packet filters[2] in the uplink TFT.
- A downlink TFT in the P-GW binds a service data flow to an EPS bearer in the downlink direction, where multiple service data flows can be multiplexed into the same EPS bearer by including multiple downlink packet filters in the downlink TFT.
- An E-RAB transports the packets of an EPS bearer between the UE and the EPC. When an E-RAB exists, there is a one-to-one mapping between the E-RAB and an EPS bearer.
- A DRB transports the packets of an EPS bearer between a UE and the eNB. When a DRB exists, there is a one-to-one mapping between this DRB and the EPS bearer/E-RAB.
- An S1 bearer transports the packets of an E-RAB between an eNB and an S-GW.
- An S5/S8 bearer transports the packets of an EPS bearer between an S-GW and a P-GW.
- A UE stores the mapping between an uplink packet filter and a DRB to create the binding between the service data flow and the DRB in the uplink.
- The P-GW stores the mapping between the downlink packet filter and S5/S8a bearer to create the binding between a service data flow and the S5/S8a bearer in the downlink.
- The eNB stores a one-to-one mapping between the DRB and the S1 bearer to create the binding between the DRB and the S1 bearer in both the uplink and downlink.
- The S-GW stores a one-to-one mapping between the S1 bearer and the S5/S8a bearer to create the binding between the S1 bearer and the S5/S8a bearer in both the uplink and downlink.

---

[2]Service data flow (SDF) is a fundamental concept in the 3GPP's definition of QoS and policy management. SDFs represent the IP packets related to a user service (web browsing, e-mail, etc.). SDFs are bound to specific bearers based on policies defined by the network operator. This binding occurs at the P-GW and the UE using traffic flow templates. TFTs contain packet filtering information to identify and map packets to specific bearers. The filters are configurable by the network operator, but at a minimum will contain five parameters, as follows:

- The source IP address.
- The destination IP address.
- The source port number.
- The destination port number.
- The protocol identification (i.e., TCP or UDP).

The policy and charging enforcement function in the P-GW filters packets.

In the downlink, the piggybacked (or encapsulated) NAS messages are used for bearer establishment, modification, or release. In the uplink, the piggybacked NAS messages are used for transferring the initial NAS messages during connection setup. The NAS messages transferred via SRB2 are also contained in RRC messages, which do not include any RRC protocol control information. Once security is established, all RRC messages on SRB1 and SRB2, including those containing NAS or non-3GPP messages, are integrity-protected and ciphered by PDCP sublayer. The NAS independently applies integrity protection and ciphering to the NAS messages.

Random access is specified entirely in the MAC sublayer including initial transmission power estimation [6]. The RRC connection establishment includes establishment of SRB1. The E-UTRAN completes the RRC connection establishment prior to completing the establishment of the S1 connection, i.e., prior to receiving the UE context from the EPC. Therefore, the AS security is not activated during the initial phase of the RRC connection. During the initial phase of the RRC connection, the E-UTRAN may configure the UE to perform measurement reporting. However, the UE only accepts handover messages when security has been activated. Upon receiving the UE context from the EPC, the E-UTRAN activates ciphering and integrity protection using the initial security activation procedure. The RRC messages to activate security are integrity-protected, while ciphering starts only after completion of the procedure, i.e., the response to the message used to activate security is not ciphered, while the subsequent messages used to establish SRB2 and DRBs are both integrity-protected and ciphered.

Following the activation of the initial security, the E-UTRAN establishes SRB2 and user DRBs, i.e., the E-UTRAN may perform this function prior to receiving the confirmation of the initial security activation from the UE. The E-UTRAN applies both ciphering and integrity protection for the RRC connection reconfiguration messages used to establish SRB2 and user DRBs. The E-UTRAN should release the RRC connection, if the initial security activation and/or the radio bearer establishment fails (i.e., security activation and user DRB establishment are triggered by a joint S1 procedure, which does not support partial success). For SRB2 and user DRBs, the security is always activated from the beginning, i.e., the E-UTRAN does not establish these bearers prior to activating security. The release of the RRC connections is initiated by the E-UTRAN. The procedure may be used to redirect the UE to another frequency or radio access technology (RAT). In certain cases, the UE may discontinue the RRC connection, i.e., it may transition to RRC_IDLE without notifying the E-UTRAN.

In the cases where the UE fails to establish a connection to the network (e.g., RLF, handover failure, RLC unrecoverable error, or reconfiguration compliance failure), the UE initiates the RRC connection reestablishment procedure, provided that security is active. If security is not active when one of the above failures occurs, the UE transitions to the RRC_IDLE state. In order to reestablish the RRC connection, the UE starts a timer known as T311 (see Section 5.6) and performs cell selection. In this case, the UE should prioritize searching of LTE frequencies.

However, no requirements are specified to prevent the UE from searching for other RATs. Upon finding a suitable cell on an LTE frequency, the UE stops timer T311, starts timer T301 and initiates a contention-based random access procedure, and sends an *RRC Connection Reestablishment Request* message. In the *RRC Connection Reestablishment Request* message, the UE includes the UE identity used in the previous cell, the identity of that cell, a short message authentication code, and a cause. The E-UTRAN uses the reestablishment procedure to continue SRB1 and to reactivate security without changing algorithms. A subsequent RRC connection reconfiguration procedure is used to resume operation on radio bearers other than SRB1 and to reactivate measurements. If the cell in which the UE initiates the reestablishment is not prepared (i.e., no UE context is available), the E-UTRAN will reject the procedure, causing the UE to transition to the RRC_IDLE state [19].

5.2.2 Quality of service control

An EPS bearer/E-RAB is the level of granularity for bearer-level QoS control in the EPC/E-UTRAN. As a result, service data flows mapped to the same EPS bearer receive the same bearer-level packet forwarding treatment, e.g., scheduling policy, queue management policy, rate shaping policy, and RLC configuration. One EPS bearer/E-RAB is established when the UE connects to a PDN and remains throughout the lifetime of the PDN connection to provide the UE with always-on IP connectivity to the PDN. This bearer is referred to as the *default* bearer. Any additional EPS bearer/E-RAB that is established to the same PDN is referred to as a *dedicated* bearer. The initial bearer-level QoS parameter values of the default bearer are assigned by the network based on subscription information. The decision to establish or modify a dedicated bearer can only be taken by the EPC, and the bearer-level QoS parameter values are always assigned by the EPC.

An EPS bearer/E-RAB is referred to as a guaranteed bit rate (GBR) bearer, if dedicated network resources related to a GBR value that is associated with the EPS bearer/E-RAB are permanently allocated by the admission control function in the eNB upon bearer establishment/modification. Otherwise, an EPS bearer/E-RAB is referred to as a non-GBR bearer. A dedicated bearer can either be a GBR or a non-GBR bearer, while a default bearer is always a non-GBR bearer. Each EPS bearer/E-RAB (GBR and non-GBR) is associated with the following bearer-level QoS parameters:

- QoS class identifier (QCI): A metric that is used as a reference to access node-specific parameters that control bearer-level packet forwarding function, e.g., scheduling weights, admission thresholds, queue management thresholds, link layer protocol configuration. These parameters are preconfigured by the network operator.
- Allocation and retention priority (ARP): The primary purpose of ARP is to decide whether a bearer establishment/modification request can be accepted or needs to be rejected in the case of resource limitation. In addition, the ARP

can be used by the eNB to decide which bearers to drop under resource shortage circumstances.
- Each GBR bearer is additionally associated with the bit rate that can be expected to be provided by that bearer.
- Each access point name (APN) access by a UE is associated with a per APN aggregate maximum bit rate (APN-AMBR).
- Each UE in the EMM-REGISTERED state is associated with a per UE-AMBR.

The GBR denotes bit rate of traffic per bearer while UE-AMBR/APN-AMBR denotes bit rate of traffic per group of bearers. Each of those QoS parameters has an uplink and a downlink component.

In Table 5.3, the delay between a policy control enforcement function (PCEF) and a radio base station (about 20 ms) should be subtracted from a given packet delay budget to derive the packet delay budget that applies to the radio interface. This delay is the average between the case where the PCEF is located close to the radio base station (approximately 10 ms) and the case where the PCEF is located far from the radio base station, e.g., in the case of roaming with home routed traffic (the one-way packet delay between Europe and the US west coast is roughly 50 ms). When calculating the average delay, it is assumed that roaming is a less typical scenario. It is expected that subtracting this average delay of 20 ms from a given packet delay budget will lead to desired end-to-end performance in most

Table 5.3 QoS Classes and Their Attributes and Usage in E-UTRA [11]

QCI	Resource Type	Priority	Packet Delay Budget (ms)	Packet Error Rate	Usage
1	GBR	2	100	10^{-2}	Conversational voice
2		4	150	10^{-3}	Conversational video (live streaming)
3		3	50	10^{-3}	Real-time gaming
4		5	300	10^{-6}	Non-conversational video (buffered streaming)
5	Non-GBR	1	100	10^{-6}	IMS signaling
6		6	300	10^{-6}	Video (buffered streaming), TCP-based (e.g., web-browsing, e-mail, chat, FTP, peer-to-peer file sharing, progressive video)
7		7	100	10^{-3}	Voice, video (live streaming), interactive gaming
8		8	300	10^{-6}	Video (buffered streaming), TCP-based (e.g., web-browsing, e-mail, chat, FTP, peer-to-peer file sharing, progressive video)
9		9			

typical cases. Note that the packet delay budget defines an upper bound. Actual packet delays for GBR traffic should typically be lower than the packet delay budget specified for a QCI as long as the UE has sufficient radio channel quality.

The rate of non-congestion related packet losses that may occur between a radio base station and a PCEF should be considered negligible. A packet error rate value specified for a standardized QCI therefore applies completely to the radio interface between a UE and radio base station.

The QCIs 1–5 and 7 are typically associated with an operator-controlled service, i.e., a service where the service data flow aggregate's uplink/downlink packet filters are known when the service data flow aggregate is authorized. In the case of E-UTRAN, this refers to the time at which a corresponding dedicated EPS bearer is established or modified.

In the case of QCI 6, if the network supports multimedia priority services, then this QCI may be used for the prioritization of non-real-time data (i.e., transport control protocol (TCP)-based services/applications) of subscribers. This QCI can be used for a dedicated premium bearer (e.g., associated with premium content) for any subscriber or a subscriber group. In this case, the service data flow aggregate's uplink/downlink packet filters are known when the service data flow aggregate is authorized. Alternatively, this QCI can be used for the default bearer of a UE/PDN for premium subscribers.

The QCI 9 is typically used for the default bearer of a UE/PDN for non-privileged subscribers. Note that AMBR can be used as a tool to provide differentiation between subscriber groups connected to the same PDN with the same QCI on the default bearer.

5.3 RRC mobility management

In order to support mobility, a UE may conduct and report to the serving eNB a series of measurements at the physical layer which include an RSRP and E-UTRA carrier received signal strength indicator (RSSI). These measurements in E-UTRA are classified as follows [4]:

- Intra-frequency E-UTRAN measurements.
- Inter-frequency E-UTRAN measurements.
- Inter-RAT measurements.

Measurement commands are used by E-UTRAN to order the UE to start measurements, modify measurements, or stop measurements. The reporting criteria that are used include event-triggered reporting, periodic reporting, and event-triggered periodic reporting. In the RRC_CONNECTED state, network-controlled UE-assisted handovers are performed and various DRX cycles are supported. In the RRC_IDLE state, cell reselections are performed and DRX is supported.

The E-UTRAN can configure the UE to perform radio signal quality measurements for handover purposes. Different measurement quantities are defined in the specification [10], and one quantity can be configured for each intra-frequency, inter-frequency, or inter-RAT measurement. The quantity configuration also defines applicable filtering, used for all reporting events associated with that quantity. The typical metrics for handover decision are as follows [8,10]:

- *Reference signal received power* (RSRP) is the linear average over the power contributions of the resource elements that carry cell-specific reference signals within the considered measurement frequency bandwidth.
- *Reference signal received quality* (RSRQ) is the ratio $N \times \text{RSRP}/(E\text{-UTRA carrier RSSI})$ where N is the number of resource blocks equivalent to the E-UTRA carrier RSSI measurement bandwidth. The measurements in the numerator and denominator are made over the same set of resource blocks.
- *UTRA FDD Common Pilot Channel (CPICH) received signal code power* (RSCP) is the received power on one code measured on the primary CPICH. The reference point for the RSCP is the antenna connector of the UE.
- *UTRA FDD carrier RSSI* is the received wideband power, including thermal noise and noise generated in the receiver, within the bandwidth defined by the receiver pulse shaping filter. The reference point for the measurement is the antenna connector of the UE.
- E_c/N_0 *UTRA FDD CPICH* is the received energy per chip divided by the power density in the band. If receiver diversity is not used by the UE, the CPICHE_c/N_0 is identical to the CPICH RSCP/UTRA carrier RSSI. The measurements are performed on the primary CPICH. The reference point for the CPICHE_c/N_0 measurement is the antenna connector of the UE.
- *GSM carrier RSSI* is the wideband received power within the relevant channel bandwidth. The measurements are conducted on a GSM BCCH carrier. The reference point for the RSSI measurement is the antenna connector of the UE.

The UE reports measurement information consistent with the measurement configuration provided by the E-UTRAN. The E-UTRAN provides the measurement configuration applicable for a UE in RRC_CONNECTED by means of dedicated signaling. The UE may be requested to perform the following types of measurements [4]:

- Intra-frequency measurements: Measurements at the downlink carrier frequency of the serving cell.
- Inter-frequency measurements: Measurements at frequencies that differ from the downlink carrier frequency of the serving cell.
- Inter-RAT measurements of UTRA frequencies.
- Inter-RAT measurements of GERAN frequencies.
- Inter-RAT measurements of CDMA2000 high rate packet data (HRPD) or 1xRTT frequencies.

The measurement configuration includes the following parameters [4]:

- *Measurement objects:* The objects on which the UE performs the measurements. For intra-frequency and inter-frequency measurements, a measurement object is a single E-UTRA carrier frequency. Associated with this carrier frequency, E-UTRAN can configure a list of cell-specific offsets and a list of blacklisted cells. The blacklisted cells are not considered in event evaluation or measurement reporting. For inter-RAT UTRA measurements, a measurement object is a set of cells on a single UTRA carrier frequency. For inter-RAT GERAN measurements, a measurement object is a set of GERAN carrier frequencies. For inter-RAT CDMA2000 measurements, a measurement object is a set of cells on a single (HRPD or 1xRTT) carrier frequency.
- *Reporting configurations*: A list of reporting configurations where each reporting configuration consists of the following [4]:
  - *Reporting criterion*: The criterion that triggers the UE to send a measurement report. This can be either periodical or a single event description.
  - *Reporting format*: The quantities that the UE includes in the measurement report and associated information (e.g., number of cells to report).
- *Measurement identities*: A list of measurement identities where each measurement identity links one measurement object with one reporting configuration. By configuring multiple measurement identities, it is possible to link more than one measurement object to the same reporting configuration, as well as to link more than one reporting configuration to the same measurement object. The measurement identity is used as a reference number in the measurement report.
- *Quantity configurations:* One quantity configuration is configured per RAT type. The quantity configuration defines the measurement quantities and associated filtering used for all event evaluation and related reporting of that measurement type. One filter can be configured per measurement quantity.
- *Measurement gaps:* Periods that the UE may use to perform measurements, i.e., no uplink or downlink transmissions are scheduled.

The E-UTRAN only configures a single measurement object for a given frequency, i.e., it is not possible to configure two or more measurement objects for the same frequency with different associated parameters, e.g., different offsets and/or blacklists. The E-UTRAN may configure multiple instances of the same event by configuring two reporting configurations with different thresholds. The UE maintains a single measurement object list, a single reporting configuration list, and a single measurement identities list. The measurement object list includes measurement objects that are specified per RAT type, including an intra-frequency object (i.e., the object corresponding to the serving frequency), inter-frequency object(s), and inter-RAT objects. Similarly, the reporting configuration list includes E-UTRA and inter-RAT reporting configurations. Any measurement object can be linked to any reporting configuration of the same RAT type. Some reporting configurations may not be linked to a measurement object.

Samsung Ex. 1005, Page 279 of 1147

Some measurement objects may not be linked to a reporting configuration. The measurement procedures distinguish the type of cells, i.e., the serving cell, listed cells (cells listed within the measurement objects), and detected cells (cells that are not listed within the measurement objects but are detected by the UE on the carrier frequencies indicated by the measurement objects).

Figure 5.7 illustrates the state transitions of a UE during normal operation. The figure shows the transition from power-on to RRC connection setup and establishment of user-plane and control-plane with the serving cell (cell selection); RRC connection release and transition to RRC_IDLE; cell reselection and transition to another cell, handover to another cell while in the RRC_CONNECTED state; and redirection from one cell to another upon RRC

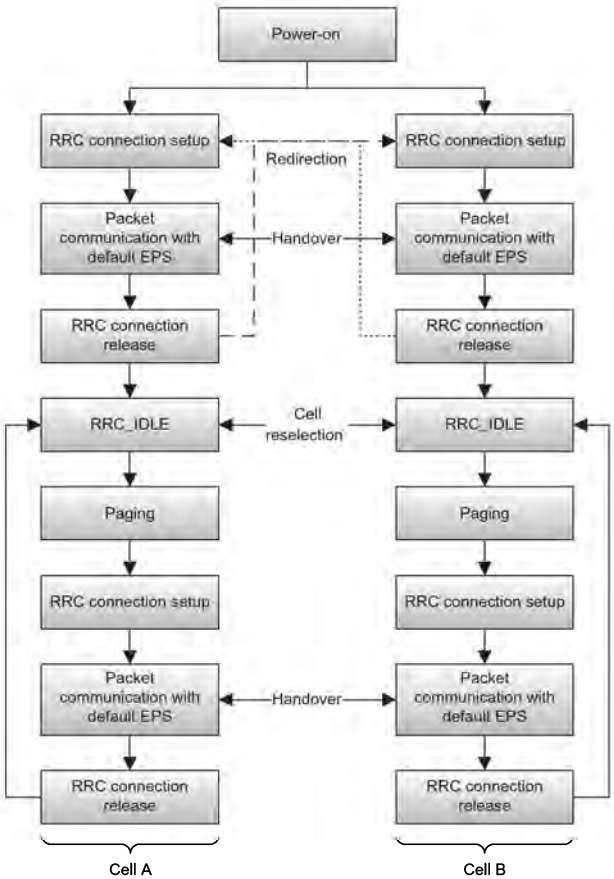


FIGURE 5.7

Overview of handover/cell selection/redirection process in E-UTRA [27].

connection release. We will explain later that cell selection, cell reselection, handover, and redirection are different processes that should not be confused.

For E-UTRA, the UE measures and reports on the serving cell, listed cells, and detected cells. For inter-RAT UTRA, the UE measures and reports on listed cells. For inter-RAT GERAN, the UE measures and reports on detected cells. For inter-RAT CDMA2000, the UE measures and reports on listed cells. For inter-RAT UTRA and CDMA2000, the UE measures and reports detected cells for the purpose of self-organizing networks. The closed subscriber group (CSG) femtocells are not included within the neighbor list. Furthermore, the assumption is that for non-femto-cell deployments, the physical cell identity is unique within the area of a large macro cell.

The intra-E-UTRAN mobility support for mobile terminals in ECM-CONNECTED handles all necessary steps for relocation or handover procedures, such as processes that precede the final handover decision on the serving network, preparation of resources on the target network, commanding the UE to the new radio resources, and finally releasing resources on the serving network side. It contains mechanisms to transfer the EU context between eNBs and to update node relations on control-plane and user-plane. The UE conducts measurements of attributes of the serving and neighbor cells to enable the process [4].

Depending on whether the UE needs transmission/reception gaps to perform the relevant measurements, measurements are classified as gap-assisted or non-gap-assisted. A non-gap-assisted measurement is a measurement on a cell that does not require transmission/reception gaps to allow the measurement to be performed. A gap-assisted measurement is a measurement on a cell that does require transmission/reception gaps to allow the measurement to be performed. Gap patterns are configured and activated by the RRC sublayer [4]. The handover procedure is performed without EPC involvement, i.e., preparation messages are directly exchanged between the eNBs. The release of the resources at the serving eNB (or source eNB) during the handover completion phase is triggered by the eNB.

The measurements to be performed by a UE for intra/inter-frequency mobility can be controlled by E-UTRAN, using broadcast or dedicated control signaling. In the RRC_IDLE state, a UE follows the measurement parameters defined for cell reselection. In the RRC_CONNECTED state, a UE follows the measurement configurations specified by RRC sublayer and are managed by the E-UTRAN. Intra-frequency and inter-frequency neighbor cell measurements are defined as follows [4]:

- Neighbor cell measurements performed by the UE are intra-frequency measurements when the current and target cell operates on the same carrier frequency.
- Neighbor cell measurements performed by the UE are inter-frequency measurements when the neighbor cell operates on a different carrier frequency compared to the current cell.

5.3.1 Mobility in RRC_CONNECTED state

In the RRC_CONNECTED state, the network controls the UE mobility, i.e., the network decides when and where the UE should initiate a handover process. For network-controlled mobility in the RRC_CONNECTED state, the network triggers the handover process based on radio conditions or network loading conditions. To facilitate this process, the network may configure the UE to perform and report measurements including the configuration of measurement gaps. The network may also initiate a blind handover, i.e., without waiting for the measurement reports from the UE. Before sending the handover message to the UE, the source eNB prepares one or more target cells. The target eNB generates the message that is used to perform the handover, i.e., the message includes the AS configuration to be used in the target cell. The source eNB transparently (i.e., does not modify values or content) forwards the handover information received from the target cell to the UE. When appropriate, the source eNB may initiate data forwarding for (a subset of) the DRBs, in order to reduce the handover interruption time.

After receiving the handover message, the UE attempts to access the target cell at the first available RACH opportunity according to random access resource selection, i.e., the handover is asynchronous. Consequently, when allocating a dedicated preamble for the random access in the target cell, E-UTRA ensures that the preamble is available from the first random access opportunity that the UE may use. Upon successful completion of the handover, the UE sends a message to confirm the handover. After the successful completion of the handover, the PDCP SDUs may be retransmitted in the target cell. This only applies to DRBs using RLC AM mode [1,4]. The source eNB should, for some time, maintain the UE context to enable the UE to return in case of handover failure. In the case of handover failure, the UE attempts to resume the RRC connection either in the source or another cell using the RRC reestablishment procedure. The connection resumption attempt succeeds only if the accessed cell is prepared, i.e., a cell belonging to the source eNB or another eNB has been prepared for the handover.

In RRC_CONNECTED state, the E-UTRAN decides to which cell a UE should hand over in order to maintain the radio link. Similar to RRC_IDLE state, the E-UTRAN may consider not only the radio link quality, but also other aspects such as UE capability, subscriber type, and access restrictions. Although the E-UTRAN may trigger handover without measurement information (blind handover), typically it configures the UE to report measurements of the candidate target cells. In LTE, the UE always connects to a single cell, thus the switching of a UE's connection from a source cell to a target cell is a hard handover. In a hard handover process, the eNB which controls the source cell requests the target eNB to prepare for the handover. The target eNB subsequently generates the RRC message to order the UE to perform the handover, and the message is transparently forwarded by the source eNB to the UE. In another type of handover supported by LTE, the UE spontaneously decides to connect to the target cell, where it then

requests that the connection be continued. The UE applies this connection reestablishment procedure only after terminating the connection to the source cell. The latter procedure only succeeds if the target cell has been prepared in advance for the handover. In addition, LTE also allows a UE to be redirected to another frequency or RAT upon connection release (see Figure 5.7). Redirection during connection establishment is not supported, since at that time the E-UTRAN may not have all the relevant information such as UE capability and the type of subscriber. However, the redirection may be performed while AS security has not been activated. When redirecting the UE to UTRAN or GERAN, the E-UTRAN may provide the system information for one or more cells on the relevant frequency. If the UE selects one of the cells for which the system information is provided, it does not need to acquire it [19].

5.3.2 Mobility in RRC_IDLE state

In the RRC_IDLE state, cell reselection between frequencies is based on absolute priorities, where each frequency has an associated priority. Cell-specific default values of the priorities are provided via the system information. In addition, E-UTRAN may assign UE-specific values upon connection release, considering factors such as UE capability or subscriber type. In case equal priorities are assigned to multiple cells, the cells are ranked based on radio link quality. Equal priorities are not applicable between frequencies of different RATs. The UE does not consider frequencies for which it does not have an associated priority. This is useful in situations where a neighboring frequency is applicable only for UEs of one of the sharing networks. Besides the cell reselection priority of a frequency, there are no idle-mode mobility-related parameters that may be assigned via dedicated signaling.

In general, the transition from RRC_IDLE to RRC_CONNECTED state is under the assumption that the RRC protection keys and user-plane protection keys are generated, while keys for NAS protection as well as higher-layer keys are assumed to be already available at the MME. On the other hand, in transitions from RRC_CONNECTED to RRC_IDLE state, the eNBs delete the stored keys and the state for the idle-mode UEs is only maintained in the MME. It is also assumed that eNB no longer stores state information about the corresponding UE and deletes the current keys from its memory.

In LTE-Advanced, the requirement for UE transition latency from the idle mode, with allocated IP address, to connected mode is less than 50 ms including the establishment of the user-plane and excluding the S1 transfer delay. The transition latency requirement from dormant sub-state (inactivity periods in short/long DRX mode) to active sub-state (continuous transmission/reception periods) in connected mode is less than 10 ms. These requirements are stricter than those for earlier releases of LTE. In the dormant sub-state, the UE has an established RRC connection and radio bearers. The UE has already been identified at cell level, but it may be in DRX mode to save power during inactivity periods. The UE may be either synchronized or unsynchronized (see Figure 5.2).

Although LTE already fulfills the latency requirements of IMT-Advanced systems, several new techniques are utilized in LTE-Advanced to further reduce the user-plane and control-plane latencies as follows [12,21]:

- Combined *RRC Connection Request* and *NAS Service Request* allow those two messages to be processed in parallel at the eNB and MME, respectively, reducing overall transition latency from idle mode to connected mode by approximately 20 ms.
- Reduced processing delays in different nodes form the major part of the delay (about 75% for the transition from idle mode to connected mode assuming a combined request); thereby any improvement has a large impact on the overall latency.
- Reduced RACH scheduling period from 10 to 5 ms results in decreasing of the average waiting time for the UE to initiate the procedure to transit from idle mode to connected mode by 2.5 ms.

Furthermore, a shorter cycle of physical uplink control channel helps reduce the average waiting time for a synchronized UE to request radio resources in the connected mode, expediting the transition from the dormant sub-state in the connected mode.

5.3.3 Intra-RAT mobility and handover

LTE supports two types of handover: X2-based handover and S1-based handover. The X2-based procedure is the least complex handover and typically would be used when an X2 interface exists between the source eNB and the target eNB. The handover is negotiated directly between the two eNBs; when the UE context is established on the target eNB, the MME is notified in order to switch the path. When no X2 interface exists between nodes, the S1-based procedure will be utilized with signaling between the target and source eNBs passing through the MME.

All handovers in LTE are hard handovers, i.e., the air-interface connection with the source eNB is disconnected before reestablishing a new connection with the target eNB. The procedure for reestablishing the connection with the target node is identical for both X2-based and S1-based handovers and requires the use of the random access procedure to synchronize and begin uplink transmissions. The UE is typically notified of the specific random access signature to use so that a contention-free procedure can be performed. From the core network perspective, handover can involve changes to the serving MME and S-GW and this can be due to mobility and the geographic location of the network nodes, or due to loading and congestion control mechanisms.

Figure 5.8 illustrates the handover procedure between eNBs within the same MME/S-GW. More information on the handover procedure can be found in Chapter 4. After the downlink path is switched at the S-GW, downlink packets on the forwarding path and on the new direct path may arrive interchanged at the

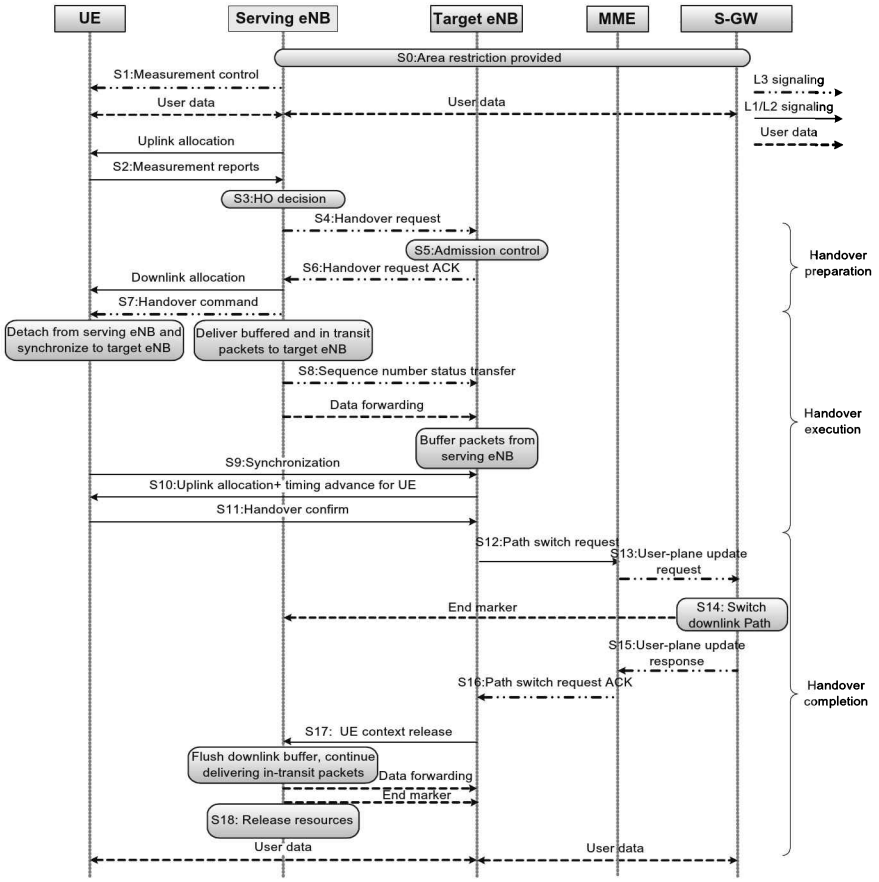


FIGURE 5.8

Intra-MME/S-GW handover procedures [4].

target eNB (see Figure 5.8). The target eNB should first deliver all forwarded packets to the UE before delivering any of the packets received on the new direct path. Upon handover, the serving eNB forwards in an orderly manner all downlink PDCP SDUs that have not been acknowledged by the UE to the target eNB. The serving eNB discards any remaining downlink RLC PDUs. Similarly, the source eNB does not forward the downlink RLC context to the target eNB.

The X2-based handover procedure can be divided into four stages (see Figure 5.8):

1. *Decision*: The E-UTRAN decides to execute a handover based on measurement reports received from the UE.
2. *Preparation*: The serving eNB notifies the target eNB that a handover is requested. It includes a list of bearers that will be transferred and whether

downlink data forwarding is necessary. The target eNB acknowledges the handover request and responds with a list of bearers that are admitted along with downlink and uplink GTP tunnel endpoints to enable data forwarding.

3. *Execution*: Once the resources have been set up on the target side, the UE is notified and detaches from the source eNB. Downlink packets that are received by the source from the S-GW are forwarded to the target eNB. PDCP status and hyperframe number information are exchanged to enable lossless handover, if required. At this point, the MME is notified of the changes in order to define direct uplink and downlink paths for the user-plane with the S-GW.
4. *Completion*: Resources in the source eNB are released following successful completion of the handover.

The message flow for an S1-based handover consists of the same four phases used for an X2-based handover. For an S1-based handover, however, there is no direct path between the eNBs; therefore, the signaling must go through the MMEs. The main stages of an S1-based handover are as follows:

1. *Decision*: The E-UTRAN decides to initiate the handover based on a measurement reports and notifies the serving MME.
2. *Preparation*: The serving MME identifies and then notifies the target MME, which begins preparation of the target resources. Because a change of S-GW is also required, the target MME also initiates preparation of resources in the target S-GW. The final preparation step is the target MME notifying the source MME of the new addresses and tunnel endpoints to enable data forwarding.
3. *Execution*: The source MME begins the execution phase of the handover by notifying the source eNB which subsequently notifies the UE. Data forwarding of downlink data begins. The UE detaches from the source eNB and synchronizes with the target eNB using random access procedures. Once established, the target MME begins the process of defining direct uplink and downlink paths for the P-GW to the target S-GW.
4. *Completion*: After successful completion of the handover, resources in the source eNB and source MME are released.

5.3.4 Inter-RAT mobility and handover

In principle, the inter-RAT handover procedures when an LTE system is the source or target are the same as the handover process in LTE. One of the differences is that upon handover to LTE, the entire AS configuration needs to be signaled, whereas within LTE it is possible to signal incremental information, where only the changes to the configuration are signaled. If ciphering was not activated in the previous RAT, the E-UTRAN activates ciphering as part of handover procedure. The E-UTRAN also establishes SRB1, SRB2, and one or more DRBs,

i.e., at least one DRB is associated with the default EPS bearer. Figure 5.9 illustrates the RRC states in E-UTRA and shows the mobility support between E-UTRAN, UTRAN, and GERAN where CCO message is used to command the UE to change to another RAT [4].

The mobility support between E-UTRAN, CDMA2000 1xRTT [17], and CDMA2000 HRPD or CDMA 1xEV-DO [13] radio access technologies are shown in Figure 5.10. In general, the procedure for handover from LTE to another RAT supports both handover and CCO with network-assisted cell change (NACC). The CCO/NACC procedure is applicable only for handover to GERAN. Handover from LTE is performed only after security has been activated. In the case of circuit-switched fallback to CDMA2000, the procedure includes support for parallel handover (i.e., to 1xRTT and HRPD), for handover to 1xRTT combined with redirection to HRPD, and for redirection to HRPD (see Figure 5.10). The UE may send a *Measurement Report* message. In the case of handover (as opposed to CCO), the source eNB requests the target node to prepare for the handover in advance. As part of the handover preparation request, the source eNB provides information about the applicable inter-RAT UE capabilities, as well as information about the currently established bearers to the target node. In response, the target node generates the handover command and returns this to the source eNB. The source eNB sends mobility from E-UTRA command to the UE, which includes the inter-RAT message received from the target cell and some inter-RAT

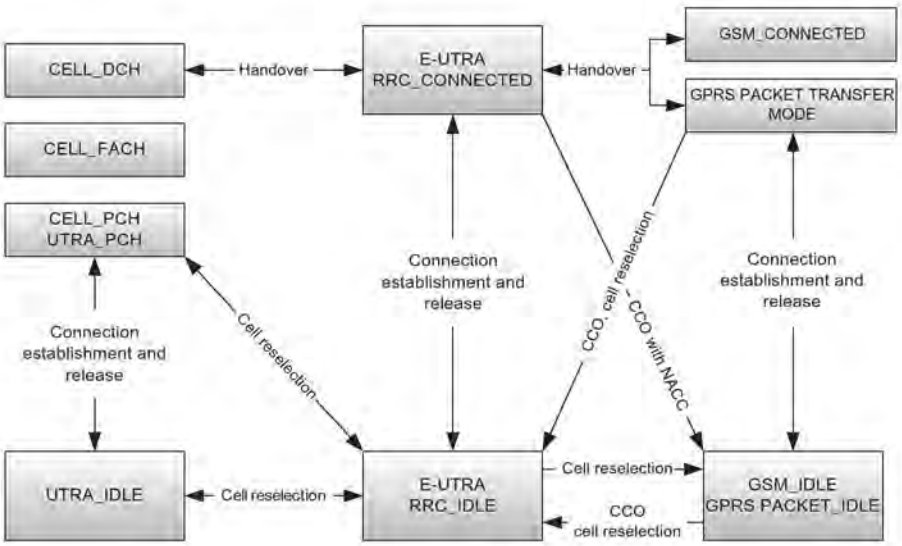


FIGURE 5.9

E-UTRA states and inter-RAT mobility procedures [1,4].

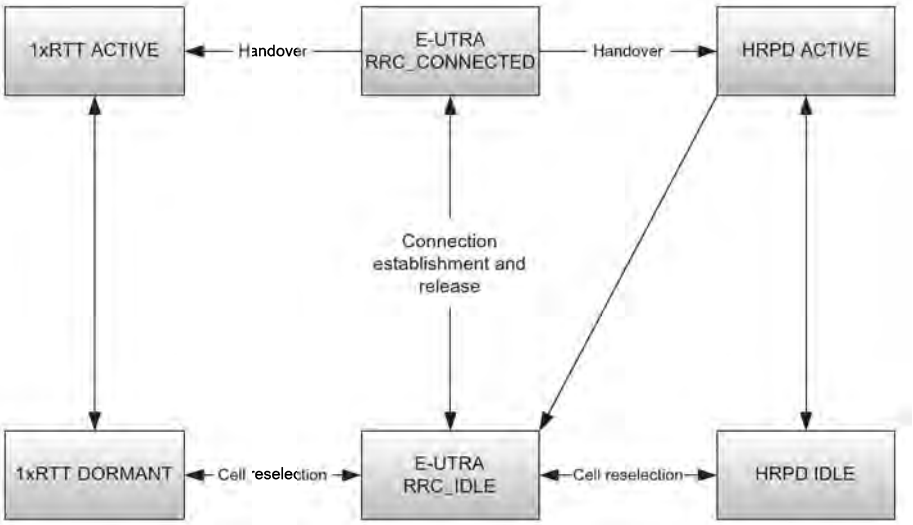


FIGURE 5.10

Mobility procedures between E-UTRA and CDMA2000 [1,4].

parameters in case of CCO. Upon receiving the message, the UE starts the timer T304 (see Section 5.6) and connects to the target node, either using the received radio configuration (handover) or by initiating connection establishment (CCO) according to the applicable specifications of the target RAT. The upper layers in the UE are informed by the AS of the bearers established by the target RAT base station. As a result, the UE is informed whether some of the previously established bearers were not admitted by the target RAT base station [19].

The inter-RAT handover procedures support signaling, conversational, and non-conversational services. The mobility between E-UTRA and non-3GPP systems other than CDMA2000 has also been studied in the literature [13–17]. In addition to the state transitions shown in Figures 5.9 and 5.10, there is support for connection release with redirection information from E-UTRA RRC_CONNECTED state to GERAN, UTRAN, and CDMA2000 (HRPD Idle/1xRTT Dormant modes).

5.4 Measurements and measurement reporting

Measurements to be performed by a UE for intra/inter-frequency mobility are controlled by the E-UTRAN, using broadcast or dedicated control signaling. In the RRC_IDLE state, a UE follows the measurement parameters defined for cell reselection by the E-UTRAN via broadcast. The use of dedicated measurement control for the RRC_IDLE state is possible through the provision of UE-specific priorities. In the RRC_CONNECTED state, a UE follows the RRC measurement

configurations defined by the E-UTRAN. Intra/inter-frequency neighbor cell measurements are defined as follows [4]:

- Intra-frequency neighbor cell measurements are performed by the UE when the current and target cell operate on the same carrier frequency. The UE conducts these measurements without measurement gaps.
- Inter-frequency neighbor cell measurements are performed by the UE when the neighbor cell operates on a different carrier frequency relative to the current cell. The UE would need measurement gaps to conduct these measurements.

The classification of the measurements to non-gap-assisted or gap-assisted depends on the UE capability and the current operating frequency. The UE determines whether a particular cell measurement needs to be performed in a transmission/reception gap, and the scheduler needs to know whether gaps are needed. The relationship between the source and target base stations' carrier frequency and bandwidth may fall into one of the following scenarios [4]:

- *Scenario A*: The current and the target cells have the same carrier frequency and cell bandwidth. This is an intra-frequency scenario without measurement gap.
- *Scenario B*: The current and the target cells have the same carrier frequency but the bandwidth of the target cell is smaller than the bandwidth of the current cell. This is an intra-frequency scenario without measurement gap.
- *Scenario C*: The current and the target cells have the same carrier frequency but the bandwidth of the target cell is larger than the bandwidth of the current cell. This represents an intra-frequency scenario without measurement gap.
- *Scenario D*: The current and the target cells have different carrier frequencies and the bandwidth of the target cell is smaller than the bandwidth of the current cell and the bandwidth of the target cell is within the bandwidth of the current cell. This is an inter-frequency scenario where measurement gap is needed.
- *Scenario E*: The current and the target cells have different carrier frequencies and the bandwidth of the target cell is larger than the bandwidth of the current cell and the bandwidth of the current cell within bandwidth of the target cell is an inter-frequency scenario where a measurement gap is needed.
- *Scenario F*: The current and the target cells have different carrier frequencies and non-overlapping bandwidths. This represents an inter-frequency scenario where a measurement gap is needed.

Measurement gap patterns are configured and activated by the RRC sublayer. When carrier aggregation is configured, the term "current cell" may refer to any serving cell of the configured set of serving cells. More specifically, the impact of carrier aggregation on the definition of intra- and inter-frequency measurements are as follows [4]:

- Intra-frequency neighbor cell measurements: Neighbor cell measurements performed by the UE are considered intra-frequency measurements when one of

the serving cells of the configured set and the target cell operate on the same carrier frequency. The UE conducts such measurements without measurement gaps.

- Inter-frequency neighbor cell measurements: Neighbor cell measurements performed by the UE are called inter-frequency measurements when the neighbor cell operates on a different carrier frequency than any serving cell of the configured set. The UE may not be able to conduct measurements without measurement gaps.

In a system with frequency reuse one, the general assumption is that mobility is mainly within the same frequency layer (i.e., between cells with the same carrier frequency). The neighbor cell measurements are needed for cells that have the same carrier frequency as the serving cell in order to ensure sustainable mobility support. The search for neighbor cells with the same carrier frequency as the serving cell, and measurements of the relevant attributes for identified cells are needed. To support mobility between different frequency layers (i.e., between cells with a different carrier frequency), the UE may need to conduct neighbor cell measurements during idle periods that are provided by DRX or packet scheduling (i.e., gap-assisted measurements).

Measurement reporting is configured as either event-triggered or periodic. Different events are defined to monitor the quality of the serving cell and neighbor cells for intra-LTE mobility (events A1 to A6) and for inter-RAT mobility (events B1and B2). For some measurements (e.g., inter-frequency), measurement gaps may be required to monitor radio quality, depending on the capabilities of the terminal. The E-UTRAN only configures a single measurement object for a given frequency, i.e., it is not possible to configure two or more measurement objects for the same frequency with different associated parameters such as different offsets and/or blacklists. Measurement gaps contain gaps every *N* LTE frames (i.e., the gap periodicity is an integer multiple of 10 ms) with 6 ms duration (Table 5.4; Figure 5.11). The triggers used by the eNB to schedule a measurement gap include the following:

- Inter-LTE mobility (event-triggered reporting criteria):
  - Event A1: Serving cell becomes better than absolute threshold.
  - Event A2: Serving cell becomes worse than absolute threshold.
  - Event A3: Neighbor cell becomes better than an offset relative to serving cell.
  - Event A4: Neighbor cell becomes better than absolute threshold.
  - Event A5: Serving cell becomes worse than one absolute threshold and neighbor cell becomes better than another absolute threshold.
  - Event A6: Neighbor cell becomes better than an offset relative to the secondary cell (SCell).
- Inter-RAT mobility (event-triggered reporting criteria):
  - Event B1: Neighbor cell becomes better than absolute threshold.
  - Event B2: Serving cell becomes worse than one absolute threshold and neighbor cell becomes better than another absolute threshold.

Table 5.4 Measurement Gap Pattern in E-UTRA [4]

Gap Pattern ID	Measurement Gap Length (ms)	Measurement Gap Repetition Period (ms)	Minimum Available Time for Inter-Frequency and Inter-RAT Measurements During 480 ms Period (ms)	Measurement Purpose
0	6	40	60	Inter-frequency E-UTRAN FDD and TDD, UTRAN FDD, GERAN, LCR TDD, HRPD, CDMA2000 1x
1	6	80	30	Inter-frequency E-UTRAN FDD and TDD, UTRAN FDD, GERAN, LCR TDD, HRPD, CDMA2000 1x

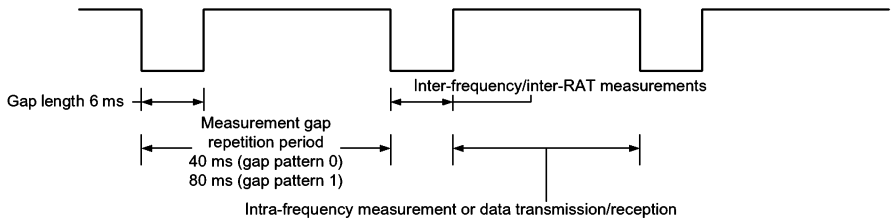


FIGURE 5.11

Illustration of measurement gap in LTE.

The UE may be configured to provide a number of periodic reports after being triggered by an event (event-triggered periodic reporting). If the UE requires measurement gaps to identify and measure inter-frequency and/or inter-RAT cells, the E-UTRAN provides a single measurement gap pattern with constant gap duration for concurrent monitoring of all frequency layers and RATs. During the measurement gaps, the UE does not transmit any data and is not expected to tune its receiver on the E-UTRAN serving carrier frequency. Inter-frequency and inter-RAT measurement requirements within this section rely on the UE being configured with one measurement gap pattern. As an example, the mobility measurement reporting for event A3 is illustrated in Figure 5.12. The figure depicts the process by which the UE reports an event A3. The UE measurements start when the RSRP of the serving cell falls below a certain level as defined by the measurement configuration parameter *s-Measure*. Event A3 is detected when the

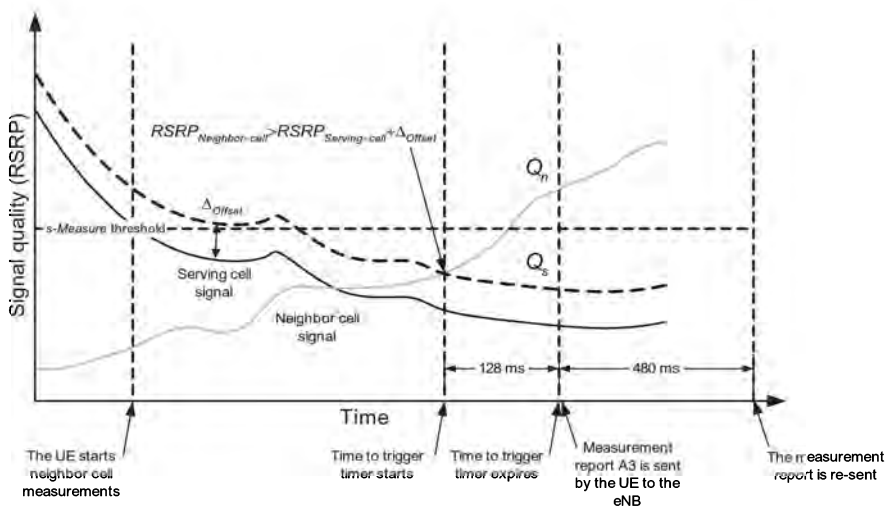


FIGURE 5.12

Illustration of mobility measurement reporting event A3 (example).

RSRP of the neighbor cell is greater than that of the serving cell plus the offset defined in the *Measurement Report Configuration* message. When this condition is met, the *Time-to-Trigger* timer is started and it expires after (for example) 128 ms. If the conditions to satisfy the event persist throughout the period that the timer is running, a measurement report is sent upon expiration of the timer. If the eNB does not respond to the initial measurement report, additional reports are sent every (for example) 480 ms, as configured [1]. Note that no hysteresis is defined for this event. The UE sends a *Measurement Report* message to inform the eNB that an event A3 has been triggered. The report includes the *Measurement Identity* which identifies the *Measurement Object*. The measured RSRPs of both the serving cell and the neighbor cell are specified in this report and correspond to the measured values at the time the report was generated, not when they were first detected. In this example, the measurement report will be re-sent every 480 ms unless the eNB responds. For the serving cell the RSRP and RSRQ is always included regardless of the triggering quantity.

To avoid UE activity outside the DRX cycle, the reporting criteria for neighbor cell measurements should match the current DRX cycle. The UE is required to apply a layer 3 filtering to each measurement quantity that it performs, which does not include quantities configured exclusively for UE receive–transmit time difference measurements. The UE is required to filter the measured result, before using for evaluation of reporting criteria or for measurement reporting, by the following formula $F_n = (1 - a)F_{n-1} + aM_n$, where M_n is the latest received measurement result from the physical layer, F_n is the updated filtered measurement result that is used for evaluation of reporting criteria or for measurement reporting, and

F_{n-1} is the previous filtered measurement result. The value of F_0 is set to M_1 when the first measurement results from the physical layer are received. The parameter a is defined as $a = 1/2^{k/4}$, in which k is the *filterCoefficent* for the corresponding measurement quantity received by the *quantityConfig*. The UE is further required to adapt the filter such that the time characteristics of the filter are preserved at different input rates, observing that the *filterCoefficentk* assumes a sample rate equal to 200 ms. If k is set to zero, no layer 3 filtering is performed. The filtering is performed in the same domain as that used for evaluation of reporting criteria or for measurement reporting, i.e., logarithmic filtering for logarithmic measurements. The filter input rate is implementation dependent in order to fulfill the performance requirements [1].

5.5 RRC procedures

The following sections describe the main RRC procedures as defined in E-UTRA.

5.5.1 System information

System information in E-UTRA is grouped into the *master information block* (MIB) and a number of *system information blocks* (SIBs) [1,4]:

- *Master information block* defines the most essential physical layer information of the cell that is required to receive other system information parts. The MIB provides the system frame number, downlink system bandwidth, and PHICH configuration.
- *System information block type 1* (SIB1) contains information for evaluating whether a UE is allowed to access a cell and further defines the scheduling of other SIBs. More specifically, SIB1 contains cell access-related information (PLMN identity list, PLMN identity, TA code, cell identity, cell status), cell selection information (minimum receiver level), and scheduling information (system information (SI) message type and periodicity, SIB mapping, SI window length).
- *System information block type 2* (SIB2) includes common and shared channel information such as access barring information (access probability factor, access class barring list, access class barring time), semi-static common channel configuration (random access parameter, PRACH configuration), and uplink frequency information (uplink EARFCN, uplink bandwidth, additional emission).
- *System information block type 3* (SIB3) comprises cell reselection information, mainly related to the serving cell. More specifically, it contains cell reselection common parameters for intra-frequency, inter-frequency and/or inter-RAT cell reselection, as well as intra-frequency cell reselection information not related to the neighbor cells.

- *System information block type 4* (SIB4) consists of information about the serving frequency and intra-frequency neighbor cells relevant to cell reselection including common and cell-specific cell reselection parameters. The information element includes cells with specific reselection parameters as well as blacklisted cells.
- *System information block type 5* (SIB5) contains information about other E-UTRA frequencies and inter-frequency neighbor cells relevant for cell reselection (including cell reselection parameters common for a frequency as well as cell-specific reselection parameters).
- *System information block type 6* (SIB6) contains information about UTRA frequencies and UTRA neighbor cells used in cell reselection including common and cell-specific cell reselection parameters.
- *System information block type 7* (SIB7) contains information about GERAN frequencies pertinent to cell reselection including cell reselection parameters for each frequency.
- *System information block type 8* (SIB8) contains information about CDMA2000 frequencies and CDMA2000 neighbor cells relevant for cell reselection including common and cell-specific cell reselection parameters.
- *System information block type 9* (SIB9) contains a home eNB name.
- *System information block type 10* (SIB10) contains an earthquake and tsunami warning system (ETWS) primary notification.
- *System information block type 11* (SIB11) contains an ETWS secondary notification.
- *System information block type 12* (SIB12) contains a commercial mobile alert service (CMAS) warning notification.
- *System information block type 13* (SIB13) contains MBMS-related information.
- *System information block type 14* (SIB14) contains information about *Extended Access Barring* for access control.
- *System information block type 15* (SIB15) contains information related to mobility procedures for MBMS reception. This information element contains the MBMS service area identities (SAI) of the current and/or neighboring carrier frequencies.
- *System information block type 16* (SIB16) contains information related to GPS time and coordinated universal time (UTC). The UE may use the parameters provided in this system information block to obtain UTC, GPS, or to obtain the local time.

The MIB is mapped to BCCH and is carried on BCH, while all other system information messages are mapped to BCCH and are dynamically carried on DL-SCH where they can be identified through system information RNTI (SI-RNTI). The MIB and SIB1 use a fixed schedule with a periodicity of 40 and 80 ms, respectively, while the scheduling of other system information messages is flexible and signaled in SIB1. The eNB may schedule DL-SCH transmissions associated with logical channels other than BCCH in the same subframe as that used

for BCCH. The paging message is used to inform the UEs in RRC_IDLE and the UEs in RRC_CONNECTED about a system information change. The system information may also be provided to the UE by means of dedicated signaling and upon handover.

Depending on the deployment, some of the SIBs may not be transmitted. For example, SIB9 is not relevant to an operator-deployed node, and SIB13 is not necessary if an MBMS service is not provided in the cell. Similar to the MIB, the SIBs are periodically broadcast. The period of an SIB transmission depends on how fast the UEs need to acquire the corresponding system information when accessing a cell. In general, lower-order SIBs are more time critical and are transmitted more often relative to higher-order SIBs. As an example, SIB1 is transmitted every 80 ms, whereas the transmission period of higher-order SIBs is flexible and can be different in different networks. The different SIBs are then mapped to different *System Information* messages, which correspond to the actual transport blocks that are transmitted on DL-SCH. SIB1 is always mapped, by itself, on to the first *System Information* message, whereas the remaining SIBs may be group-wise multiplexed on to the same SI subject to the following constraints: (1) the SIBs are mapped to the same *System Information* message, if they have the same transmission period, e.g., two SIBs with a transmission period of 320 ms can be mapped to the same *System Information* message, whereas an SIB with a transmission period of 160 ms must be mapped to a different *System Information* message; (2) the total number of information bits that is mapped to a single *System Information* message must not exceed the permissible size of the transport block. It should be noted that the transmission period of a given SIB may be different in different networks. For example, different operators may have different requirements regarding the period when different types of neighbor cell information is transmitted. Furthermore, the amount of information that can fit into a single transport block depends on the deployment parameters such as cell bandwidth and cell size [20].

The change of system information (other than ETWS) only occurs at specific radio frames. The system information may be transmitted a number of times with the same content within a modification period as defined by its scheduling. When the network changes some of the system information, it notifies the UEs about this change. In the next modification period, the network transmits the updated system information, as illustrated in Figure 5.13 where different colors indicate

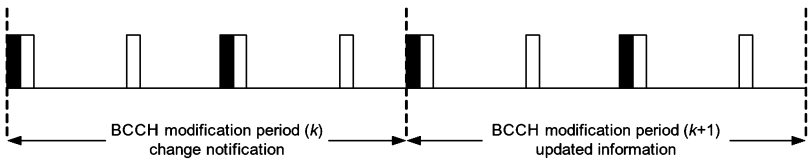


FIGURE 5.13

Change of system information timing [4].

different system information. Upon receiving a change notification, the UE acquires the new system information immediately from the start of the next modification period. The UE applies the previously acquired system information until the UE acquires the new system information.

The paging message is used to inform UEs in RRC_IDLE and UEs in RRC_CONNECTED about a system-information change. If the UE receives a paging message indicating a system information modification, the UE by default expects that the system information will change at the next modification period boundary (see Figure 5.13). Although the UE may be informed about changes in system information, no further details such as which system information is changing are provided. The *SIB1* includes a value tag that indicates if a change has occurred in the *System Information* messages. The UEs may use the tag upon return from out-of-coverage to verify whether the previously stored *System Information* messages are still valid. Furthermore, the UE considers stored system information to be invalid after 3 hours from the moment it was successfully confirmed as valid, unless otherwise specified.

The E-UTRAN may not update the tags upon change of some system information such as ETWS information, or regularly changing parameters such as CDMA2000 system time. Similarly, the E-UTRAN may not include the system information modification within the paging message upon change of some system information. The UE verifies that stored system information remains valid by either checking the tags in *SIB1* after the modification period boundary or by attempting to find the system information modification indication during the modification period in case no paging is received in every modification period. If no paging message is received by the UE during a modification period, the UE may assume that no change of system information will occur at the next modification period boundary. If a UE in RRC_CONNECTED state, during a modification period, receives one paging message, it may assume from the presence/absence of system information modification whether a change of system information other than ETWS information will occur in the next modification period.

The UE performs the system information acquisition procedure to acquire the AS and NAS system information that is broadcast by the E-UTRAN. The procedure applies to UEs in RRC_IDLE and UEs in RRC_CONNECTED. The UE applies the system information acquisition procedure upon cell selection and reselection, after handover completion, after entering E-UTRA from another RAT, upon return from out-of-coverage, upon receiving a notification that the system information has changed, upon receiving an indication about the presence of an ETWS notification, upon receiving a request from CDMA2000 upper layers, and upon exceeding the maximum validity duration. Unless explicitly stated in the procedural specification, the system information acquisition procedure overwrites any stored system information [1].

If the UE is in the RRC_IDLE state, it must ensure that it has a valid version of the *MIB* and *SIB1*, as well as *SIB2* through *SIB8*, depending on the support of other radio access technologies. If the UE is in the RRC_CONNECTED state, it

must ensure that it has the *MIB*, *SIB1*, and *SIB2* as well as *SIB8*, depending on the UE's support of CDMA2000 [18].

In LTE, the system information is classified into five categories as follows: (1) information valid across multiple cells; (2) information needed at cell search; (3) information needed prior to cell camping; (4) information needed before cell access; and (5) information needed while camping on a cell. From the UE perspective, the information that is needed at cell selection and prior to camping are very similar. Before a UE can camp on a cell, it needs to know if the access is allowed to that cell. Thus, it would be very beneficial to know all access restrictions already at cell search phase. In order to support full mobility within the serving frequency layer, the UEs need to perform a cell search periodically, and thus it is important that the information needed for a cell search is readily available to reduce cell search latency and minimize UE power consumption.

More specifically, before a UE can camp on a cell, it needs to know any access-related parameters in order to avoid camping on cells where access is restricted, i.e., any cell access restriction parameters such as TA identity, cell barring status, and cell reservation status. Thus, the UE needs to know whether the cell is barred or reserved in order to avoid camping on a barred cell. Also barring time might be needed in order to ensure that the UE does not have to poll barring time frequently from the system information. Another option is that barring status is indicated also in the neighbor cell list and radio access limitation parameters, i.e., any radio condition parameters that limit the access to the cell [1].

When a UE has camped on a cell, it needs to continue measuring the neighbor cells in order to stay camped. In order for the UE to start mobility procedures, it needs to receive parameters such as reporting periods, reporting event parameters, and time to trigger. The UEs in the RRC_IDLE state need cell reselection parameters. The UEs in the RRC_CONNECTED state need parameters of the neighbor cells for handover and for error recovery cases. Neighbor cell lists are needed to start neighbor cell measurements. UEs in different states may use different sets of neighbor cell lists. The LTE system information can be classified into two distinctive groups in terms of timing: static and flexible. The static part is sent more frequently, e.g., once per frame, in the cell and has a limited capacity for information transfer. The flexible part has a flexible amount of scheduled resources available and thus encompasses most of the system information. The flexible part has different types of information elements which require independent scheduling in order to allow sufficiently fast reception and efficient resource utilization.

5.5.2 Paging

The DRX reduces UE power consumption in the idle mode. The UE monitors only one paging occasion (PO) every DRX cycle. This is the shortest of the UE-specific DRX cycle and the default DRX cycle broadcast in SIB2. The default

Table 5.5 UE and Network Identifiers in LTE/LTE-Advanced [27]

Abbreviation	Identifier	Description	Value/Structure
IMSI	International mobile subscriber identity	Unique identification of mobile subscriber in the network	IMSI (not more than 15 digits) = PLMN ID + MSIN = MCC + MNC + MSIN
PLMN ID	PLMN identifier	Unique identification of PLMN	PLMN ID (not more than 6 digits) = MCC + MNC
MCC	Mobile country code	Assigned by ITU	3 digits
MNC	Mobile network code	Assigned by national authority	2 or 3 digits
MSIN	Mobile subscriber identification number	Assigned by operator	9 or 10 digits
GUTI	Globally unique temporary UE identity	Identifies a UE (between the UE and the MME) on behalf of IMSI for security reasons	GUTI (not more than 80 bits) = GUMMEI + M-TMSI
TIN	Temporary identity used in next update	GUTI is stored in TIN parameter of UE's context. TIN indicates which Temporary ID will be used in the next update	TIN = GUTI
S-TMSI	SAE temporary mobile subscriber identity	To locally identify a UE in short form within an MME group (unique within an MME Pool)	S-TMSI (40 bits) = MMEC + M-TMSI
M-TMSI	MME mobile subscriber identity	Unique within an MME	32 bits
GUMMEI	Globally unique MME identity	To identify an MME uniquely in a global sense	GUMMEI (not more than 48 bits) = PLMN ID + MMEI
MMEI	MME identifier	To identify an MME uniquely within a PLMN	MMEI (24 bits) = MMEGI + MMEC

(Continued)

Table 5.5 (Continued)

Abbreviation	Identifier	Description	Value/Structure
MMEGI	MME group identifier	Unique identification of an MME group within a PLMN	16 bits
MMEC	MME code	To identify an MME uniquely within an MME Group	8 bits
C-RNTI	Cell-radio network temporary identifier	To identify a UE uniquely in a cell	0 × 0001 ~ 0xFFF3 (16 bits)
IMEI	International mobile equipment identity	To identify an ME (Mobile Equipment) uniquely	IMEI (15 digits) = TAC + SNR + CD
IMEI/SV	IMEI/ software version	To identify an ME (Mobile Equipment) uniquely	IMEI/SV (16 digits) = TAC + SNR + SVN
ECGI	E-UTRAN cell global identifier	To identify a cell in global (Globally Unique) EPC can know UE location based on ECGI	ECGI (not more than 52 bits) = PLMN I D + ECI
ECI	E-UTRAN cell identifier	To identify a cell within a PLMN	ECI (28 bits) = eNB ID + cell ID
PGW ID	PDN GW identity	To identify a specific PDN GW (P-GW) HSS assigns P-GW for PDN (IP network) connection of each UE	IP address (4 bytes) or FQDN (variable length)
TAI	Tracking area identity	To identify TA which is globally unique	TAI (not more than 32 bits) = PLMN ID + TAC P-GW
TAC	Tracking area code	To indicate to which TA the eNB belongs (per cell) which is unique within a PLMN	16 bits

(Continued)

Table 5.5 (Continued)

Abbreviation	Identifier	Description	Value/Structure
TAI list	Tracking area identity list	UE can move into the cells included in TAI list without location update (TA update) which is globally unique	Variable length
PDN ID	Packet data network identity	To identify a PDN (IP network) with which mobile data user communicates PDN Identity (APN) is used to determine the P-GW and point of interconnection with a PDN With APN as query parameter to the DNS procedures, the MME will receive a list of candidate P-GWs, and then a P-GW is selected by MME	PDN Identity = APN = APN−NI + APN−OI (variable length) APN−NI: APN Network Identifier APN−OI: APN Operator Identifier
EPS bearer ID	Evolved packet system bearer Identifier	To identify an EPS bearer (Default or Dedicated) per UE	4 bits
E-RAB ID	E-UTRAN radio access bearer identifier	To identify an E-RAB per UE	4 bits
DRB ID	Data radio bearer identifier	To identify a DRB per UE	4 bits
LBI	Linked EPS bearer ID	To identify the default bearer associated with a dedicated EPS bearer	4 bits

(*Continued*)

Table 5.5 (Continued)

Abbreviation	Identifier	Description	Value/Structure
TEID	Tunnel endpoint Identifier	To identify the endpoint of a GTP tunnel when the tunnel is established	32 bits

Table 5.6 E-UTRA Paging Parameters [2]

	FDD				TDD			
N_s	PO when $i\_s=0$	PO when $i\_s=1$	PO when $i\_s=2$	PO when $i\_s=3$	PO when $i\_s=0$	PO when $i\_s=1$	PO when $i\_s=2$	PO when $i\_s=3$
1	9	N/A	N/A	N/A	0	N/A	N/A	N/A
2	4	9	N/A	N/A	0	5	N/A	N/A
4	0	4	5	9	0	1	5	6

paging cycle can be 320, 640, 1280, or 2560 ms. The PO is defined in terms of paging frame (PF) and subframe index. Every UE uses its own IMSI (Table 5.5) to determine the PF, and within the PF, the PO. The network uses the S-TMSI, which is part of the globally unique temporary identifier (GUTI), to send the paging message to the UE. The GUTI is assigned to the UE during the attach procedure. If the network uses the IMSI for paging, the UE locally detaches and initiates the attach procedure.

One PO is a subframe containing a PDCCH, whose CRC is scrambled with P-RNTI, for the purpose of addressing the paging message. One PF is one radio frame, which may contain one or multiple PO(s). When idle-mode DRX is enabled, the UE needs only to monitor one PO per DRX cycle [2]. The parameters PF and PO are determined by following equations using the DRX parameters provided in the *System Information* message:

- PF is given by following equation: SFN mod $T = (T/N)$(UE_ID mod N).
- i_s is an index which points to a PO (Table 5.6) and is derived from the following equation: $i\_s = \lfloor \text{UE\_ID}/N \rfloor \text{mod} N_s$.
- $T = \min(T_{\text{UE-specific-paging-cycle}}, T_{\text{Cell-specific-paging-cycle}})$.

The DRX parameters stored in the UE must be updated whenever the DRX parameter values are changed in the system information. If the UE has no IMSI, for instance, when making an emergency call without USIM, the UE uses, by default, UE_ID = 0 in the above equations. In the above equations, T denotes the DRX cycle of the UE. The value of T is determined by the smallest of the

UE-specific DRX values, if allocated by upper layers, and a default cell-specific DRX value is broadcast in the system information. If UE-specific DRX is not configured by upper layers, the default value is applied; $nB = \{4T, 2T, T, T/2, T/4, T/8, T/16, T/32\}$ is used as one of parameters to derive the PF and PO and is broadcast in SIB2; $N = \min(T, nB)$indicates the number of PFs within the paging cycle of a UE; $N_s = \max(1, nB/T)$(i.e., number of POs (subframes) per PF); and UE_ID = IMSI mod 1024 is the same as the UE Identity Index provided to the eNB by the MME through the S1-AP *Paging* message. Note that IMSI is given as sequence of digits of type integer and in the above equation is interpreted as a decimal integer number, where the first digit given in the sequence represents the highest order digit [2]. Table 5.6 provides the relationship between the paging parameters in FDD and TDD (for all downlink/uplink ratios) systems.

As we just learned, in the idle mode, the UE does not have to continuously monitor the paging channel. Based on the paging control channel configuration information that is broadcast in SIB2, the UE can go to the DRX mode between POs. Let us try to understand this process through an illustrative example. Figure 5.14 illustrates an example where the paging cycle defined by the *defaultPagingCycle* parameter is set to $T = T_{\text{UE-specific-paging-cycle}} = T_{\text{Cell-specific-paging-cycle}} = 128$ frames. The total number of POs per paging cycle is defined by parameter *nB*, which is broadcast in SIB2. In this example (assuming an FDD system), we assume that:

$nB = T; \quad N = 128, N_s = 1$
IMSI of UE (in decimal) = 262022008880715
UE_id = 262022008880715 mod 1024 = 587
$i\_s = 0 \rightarrow \text{PO} = 9$

Thus, one PO is available in every paging cycle. The specific SFN number that each UE should monitor during each paging cycle is derived from the UE's IMSI. The maximum number of POs that can be defined per radio frame is 4 ($nB = 4T$), and these will only be transmitted in subframes 0, 4, 5, and 9 (see Table 5.6). Therefore, in this example, whenever SFN mod 128 = 75, the UE will wake up and look for a possible paging message addressed to it in the 9th subframe of PF (i.e., SFN = 75, 203, 331, ...) The value of N_s is cell-specific and is set according to the cell load and paging capacity. The choice of 4th and 9th subframes for lower paging capacity configurations is due to the fact that these subframes are adjacent to the 5th and 0th subframes which carry the primary and secondary synchronization signals. This allows the UE to detect the synchronization signals at the same time that it wakes up to check for a paging message. The 0th and the 5th subframes are not selected for narrowband configurations because the 0th subframe carries the MIB and the 5th subframe includes SIB1. For wideband systems, there are usually sufficient resources to accommodate both MIB/SIB1 and the paging messages.

We now change of our focus to tracking the UEs in the RRC_IDLE mode and how paging messages are sent to these UEs in the network. The MME keeps track

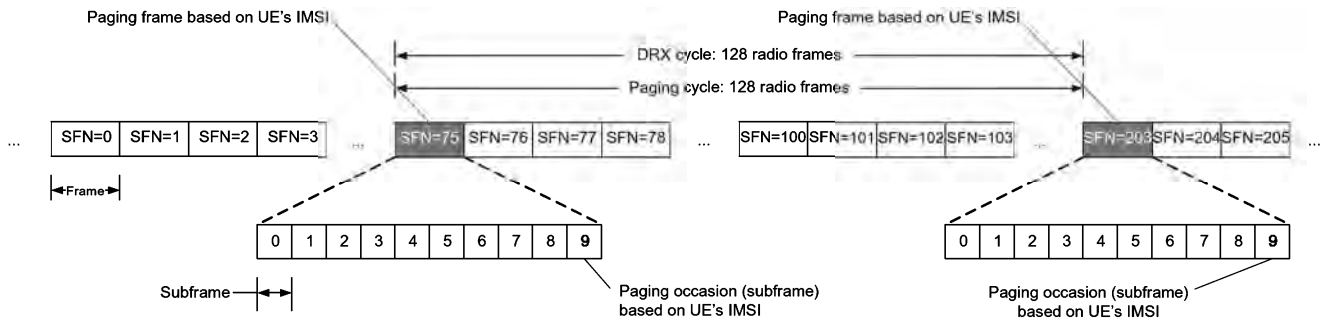


FIGURE 5.14

Illustration of an example paging procedure.

of the location of all UEs in its service area. When a UE registers in the network, the MME creates an entry for the UE and signals the location to the home subscriber server (HSS) in the UE's home network. The MME requests the appropriate resources to be set up in the eNB, as well as in the S-GW which it selects for the UE. The MME then continues tracking the UE's location either at the level of eNB, if the UE remains connected or at the level of a tracking area, which is a group of eNBs which send the paging messages to the UEs in the idle mode (Figure 5.15). The MME controls the setup and release of resources based on the UE's activity. The MME also participates in control signaling for handover of an active mode UE between eNBs, S-GWs, or other MMEs. The MME is involved in every eNB change. An idle-mode UE reports its location either periodically or when it moves to another TA. If data is received from the external networks for an idle-mode UE, the MME will be notified, and it subsequently requests the eNBs in the UE's TA to page the UE [22].

The UE's location in the RRC_IDLE state is known by the MME with the granularity of a TA. The size of the TA can be optimized in network planning

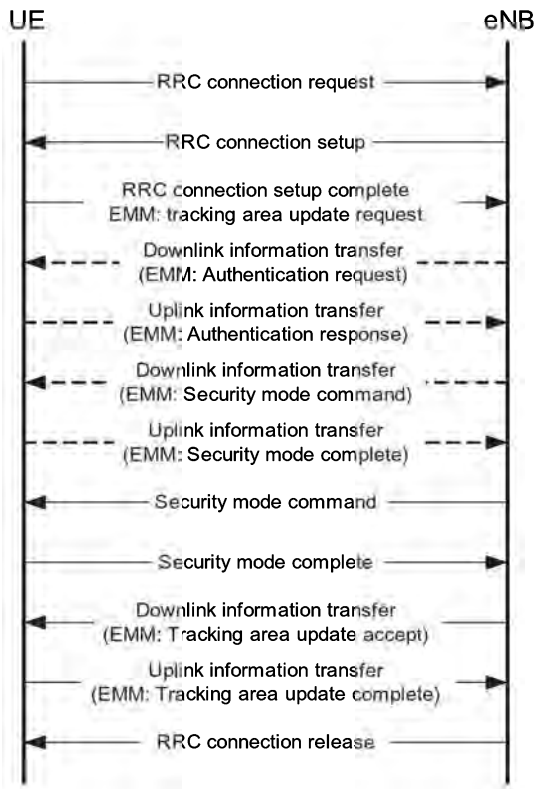


FIGURE 5.15

Tracking Area Update procedure in E-UTRA [1].

phase. A larger TA is beneficial to avoid frequent TA update signaling. On the other hand, a smaller TA is beneficial to reduce the paging signaling load for the incoming packet calls. The UE can be assigned multiple TAs to avoid unnecessary TA updates at the TA boundaries, e.g., when the UE moves back and forth between cells of two different TAs. The UE can also be assigned both an LTE TA and a UTRAN routing area to avoid signaling when changing between the two systems [22].

When an idle-mode UE moves across a TA boundary, this event will trigger the TA update procedure, which is initiated by the UE (Figure 5.16). In this case, the UE sends to the (new) target resources a *Tracking Area Update Request* message. The target MME requests and receives the bearer context information from the (old) source MME. Authentication may be performed at this point, if the source MME does not authenticate the UE. After authentication, the target MME sends a *Create Session Request* message to the target S-GW, which triggers a *Modify Bearer Request* message to be sent to the P-GW. The P-GW responds to the S-GW with a *Modify Bearer Response* message which, in turn, responds to the MME with a *Create Session Response* message. Then, the location of the UE needs to be updated in the HSS. This is accomplished by the target MME sending an *Update Location Request* to the HSS, which cancels the location of the UE in the source MME and confirms the location update to the target MME. Finally, the *Tracking Area Update* is confirmed with a *Tracking Area Update Complete* NAS PDU sent as part of a

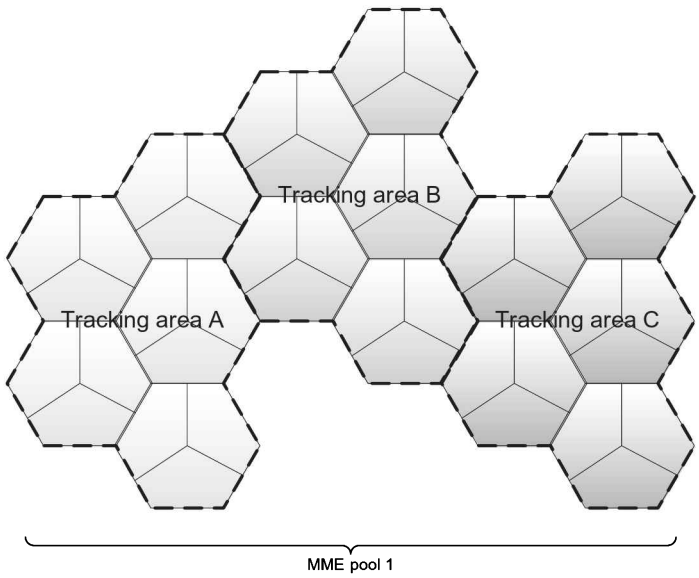


FIGURE 5.16

Illustration of TAs and MME pool [18,22].

downlink information transfer message from the target eNB to the UE. The UE initiates the *Tracking Area Update* procedure in two scenarios:

- When the TA has changed (the UE reselects a cell that does not belong to the TAI list provided by the MME in the *Attach Accept*). The TAI is used to identify TAs. The TAI is constructed from the PLMN identity which the TA belongs to and the tracking area code (TAC) of the TA.
- When timer T3412 expires (also provided in the *Attach Accept* or the last *Tracking Area Update Accept*).

In EPS and during normal operation, the UE registers with an MME group, which may have multiple MMEs sharing the traffic load (see Figure 5.16). This group may have one or more TAs. This implies that the UE can roam in the coverage areas of the eNBs broadcasting these TAs without having to re-register with the network. Note that a UE in the ECM-IDLE (or RRC_IDLE) state is paged in all cells of the TAs in the TAI list. When the UE moves into an eNB coverage area that broadcasts a different TAI, registration with the network is required. In the context of EPS, this is known as the TA update procedure. An EMM-REGISTERED UE also performs periodic TA updates with the network after the periodic TAU timer expires. The UE is still covered by the eNB that is broadcasting the TAC assigned by the registering MME. If GUTI is allocated or changed, the *Tracking Area Update Complete* message must be sent back to the eNB.

At this point, since we have referred to various UE and network identifiers in this chapter, it is worthwhile to review the definition, construction, and application of those identifiers to help better understanding of the procedures described in this chapter. The classification and structure of those identifiers are depicted in Figure 5.17 and summarized in Table 5.5.

5.5.3 PLMN and cell selection

Before the UE can camp on a cell, the cell must meet a number of requirements. Depending on the UE's subscription or access class, the UE may or may not have access to certain cells due to access barring, closed-subscriber-group femto-cell, or reserved cell restrictions. The UE measures the RSRP based on cell-specific reference signals to determine whether the candidate cell satisfies a certain criterion S. The criterion S is also considered during cell reselection. The cell selection criterion S is satisfied when the following conditions are met [2]:

$$S_{\text{rxlev}} > 0 \quad S_{\text{qual}} > 0$$

where

$$S_{\text{rxlev}} = Q_{\text{rxlevmeas}} - (Q_{\text{rxlevmin}} + Q_{\text{rxlevminoffset}}) - P_{\text{compensation}}$$

$$S_{\text{qual}} = Q_{\text{qualmeas}} - (Q_{\text{qualmin}} + Q_{\text{qualminoffset}})$$

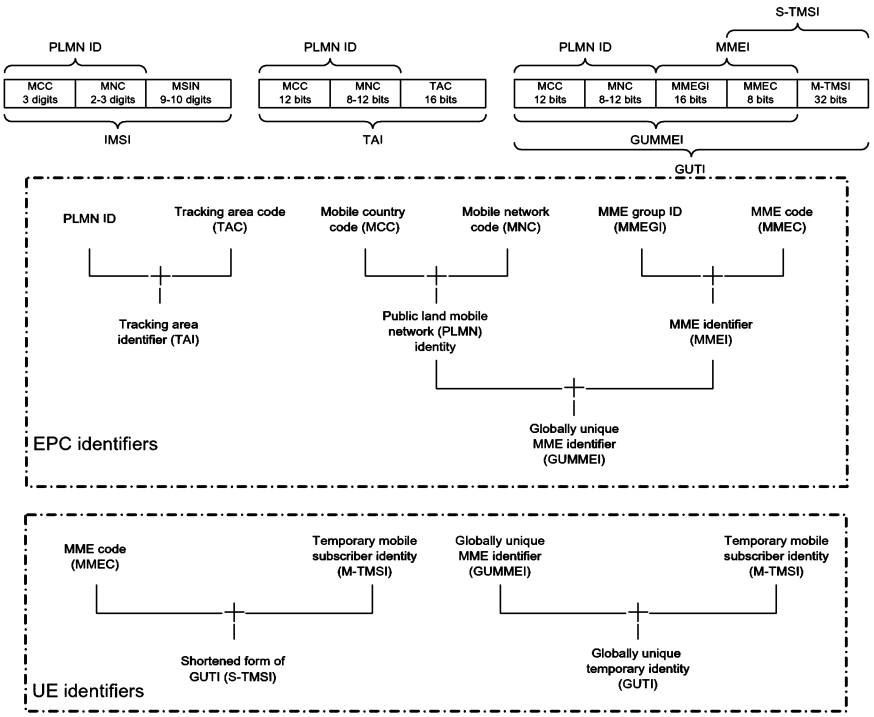


FIGURE 5.17

Classification of UE and network identifiers in LTE/LTE-Advanced [26,27].

In the above equations

- S_{rxlev}: Cell selection receive level value (dB).
- S_{qual}: Cell selection quality value (dB).
- $Q_{\text{rxlevmeas}}$: Measured cell receive level value (RSRP).
- Q_{qualmeas}: Measured cell quality value (RSRQ).
- Q_{rxlevmin}: Minimum required receive level in the cell (dBm).
- Q_{qualmin}: Minimum required quality level in the cell (dB).
- $Q_{\text{rxlevminoffset}}$: Offset to the signaled Q_{rxlevmin} taken into account in the S_{rxlev} evaluation as a result of a periodic search for a higher priority PLMN while camped normally in a visited PLMN.
- $Q_{\text{qualminoffset}}$: Offset to the signaled Q_{qualmin} taken into account in the S_{qual} evaluation as a result of a periodic search for a higher priority PLMN while camped normally in a visited PLMN.
- $P_{\text{compensation}} = \max(P_{\text{Emax}} - P_{\text{PowerClass}}, 0)$ (dB).
- P_{Emax}: Maximum transmit power level that the UE may use when transmitting in the uplink in the cell (dBm) as defined in 3GPP TS 36.101 [8].

- $P_{PowerClass}$: Maximum RF output power of the UE (dBm) according to the UE power class as defined in 3GPP TS 36.101 [8].

The signaled values $Q_{rxlevminoffset}$ and $Q_{qualminoffset}$ are only applied when a cell is evaluated for cell selection as a result of a periodic search for a higher priority PLMN while camped normally in a visited PLMN. During this periodic search for a higher priority PLMN, the UE may check the criterion S of a cell using parameter values stored from a different cell of this higher priority PLMN.

After the UE is camped for normal service, it performs attach and PDN connectivity procedures to register its location with the selected PLMN and obtain TA information. This enables the PLMN to reach the UE. In addition, the UE is required to measure the RSRP level of the serving cell and evaluate the cell selection criterion S defined for the serving cell at least every DRX cycle. The UE is further required to identify new intra-frequency cells and perform RSRP measurements on the identified intra-frequency cells without an explicit intra-frequency neighbor list containing physical cell identities. The LTE specifies the time duration for a UE to detect a cell, as well as the frequency at which the UE measures RSRP. This is true for cells on the serving frequency or other frequencies. The standard also specifies the RSRP levels that are used for the detection of the cells. The UE is required to apply filtering on RSRP during measurements. The NAS is informed, if cell selection and reselection results in changes in the received system information relevant to NAS.

SIB1, SIB3, and SIB4 provide parameters that are used during cell reselection. SIB3 contains measurement rule parameters. The E-UTRAN requires the UEs to detect and measure neighbor cells when the serving cell S_{rxlev} is lower than a specific threshold. The UE may choose not to perform the measurements, if the serving cell's metric is greater than the threshold. After conducting the measurements, the UE applies the ranking criterion and the highest ranked cell is selected during cell reselection. Note that hysteresis is applied through $Q_{Hysteresis}$ and Q_{offset} before the new cell is selected. During the DRX cycle, if the UE has determined that the serving cell does not fulfill the cell selection criterion S, it initiates measurements of all neighbor cells, irrespective of the measurement rules currently limiting UE measurement activities. If a UE in the RRC_IDLE state searches for 10 seconds without finding a new suitable cell based on searches and measurements using the intra-frequency, inter-frequency, and inter-RAT information indicated in the *System Information*, the UE is considered to be out of service area.

Figure 5.18 demonstrates an intra-frequency cell reselection procedure. In this figure, the cell-ranking criterion R_s for a serving cell and R_n for neighbor cells is defined by [2]:

$$\begin{aligned} R_s &= Q_{meas(s)} + Q_{Hysteresis} \\ R_n &= Q_{meas(n)} - Q_{offset} \end{aligned}$$

- Q_{meas}: RSRP measurement quantity used in cell reselections.

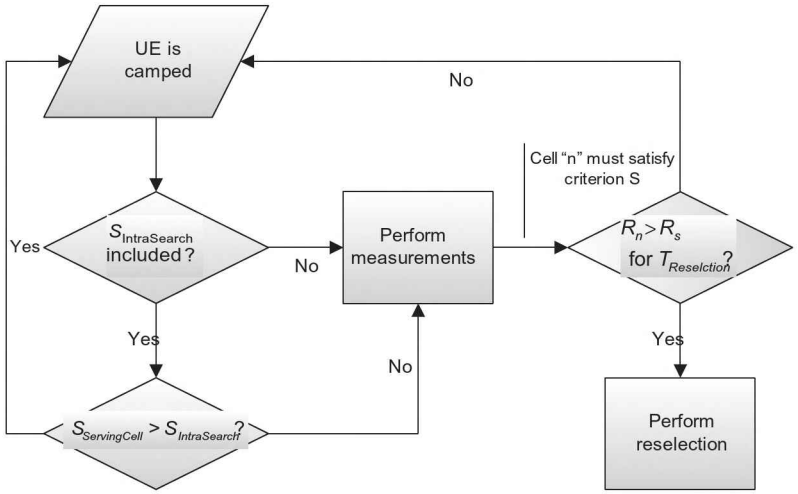


FIGURE 5.18

Illustration of intra-frequency cell reselection [1].

- Q_{offset}: For intra-frequency: $Q_{offset} = Q_{offset(s,n)}$ if $Q_{offset(s,n)}$ (the offset between the two cells) is valid; otherwise $Q_{offset} = 0$. For inter-frequency: $Q_{offset} = Q_{offset(s,n)} + Q_{offset_{frequency}}$ if $Q_{offset(s,n)}$ is valid; otherwise, $Q_{offset} = Q_{offset_{frequency}}$ (frequency-specific offset for equal priority E-UTRAN frequencies).
- $Q_{Hysteresis}$: Hysteresis value for ranking criteria.

After a UE is powered-up and a PLMN is selected, the cell selection process is initiated. This process allows the UE to select a suitable cell to camp on in order to access network services. During this process, the UE can use initial cell selection or stored information for cell selection. Initial cell selection requires no prior knowledge of E-UTRAN RF channels and the UE is required to scan all RF channels in the E-UTRA bands according to its capabilities to find a suitable cell [8,9]. On each carrier frequency, the UE only searches for the strongest cell. Once a suitable cell is found, the cell is selected. Cell selection based on stored information utilizes the UE stored information of carrier frequencies and cell parameters from previously received measurement control information elements or from previously detected cells. If the UE cannot find any suitable cell based on the latter approach, the initial cell selection procedure is performed. The UE normally operates in its home PLMN. A visited PLMN is selected when the coverage is lost. The visited PLMN may also be manually selected by the user.

5.5.4 Cell reselection

Inter-frequency cell reselection is based on absolute priorities that are provided in SIB5 and are valid for all UEs in the serving cell. In addition, UE-specific

priorities may be signaled as part of the *RRC Connection Release* message. The dedicated signaling priorities are deleted when the UE enters the RRC_CONNECTED state, timer T320 expires, or when PLMN selection is performed. Only the frequencies listed in SIB5 are considered for inter-frequency reselection. The list can contain a maximum of eight inter-frequencies that the UE may be allowed to monitor in E-UTRAN. The parameters provided in SIB3 are also considered for ranking evaluations.

Once a cell is selected or reselected, the UE is not allowed to reselect a new cell for at least one second. Furthermore, the latter constraint ensures that reselection decisions are not made in response to instantaneous RF signal variations. For inter-frequency neighbor cells, it is possible to indicate a cell-specific offset to be considered during reselection. These parameters are common to all cells on a different frequency. Blacklists can be provided to prevent the UE from reselecting the specific intra/inter-frequencies. Cell reselection can also be mobility dependent (based on UE speed). SIB6, SIB7, and SIB8 provide the related information for UTRAN/GERAN/CDMA2000, respectively.

If the $S_{\text{ServingCell}}$ of the E-UTRA serving cell (or other cells on the same frequency layer) is greater than $S_{\text{nonIntraSearch}}$, the UE may not search or measure inter-RAT frequency layers of equal or lower priority. If the UE measures the UTRA FDD cells, it is required to measure CPICH E_c/I_0 and CPICH RSCP of detected UTRA FDD cells in the neighbor cell list. Hysteresis, thresholds, suitability, and priorities are considered during the reselection decision to a UTRA cell. For GSM, the signal level of the GSM BCCH carrier of each GSM neighbor cell indicated in the measurement control system information of the serving cell is measured. For HRPD (1xRTT), the UE is required to measure CDMA2000 HRPD pilot strength of HRPD cells (CDMA2000 1x RTT pilot strength of CDMA2000 1X cells) in the neighbor cell list. In addition, the UE may be requested to pre-register with the HRPD network, where the network identifies the pre-registration zone for the UE. Pre-registration reduces the time required for cell reselection and handover by enabling HRPD and point-to-point protocol (PPP) session establishment in advance. The E-UTRA node acts as a relay agent to transfer 1xEV-DO *Session Setup* messaging.

As mentioned earlier, after camping on a suitable E-UTRAN cell, the UE uses a NAS registration procedure to register its presence in the TA of the selected cell. This allows the PLMN to page the user in the registered TA(s). If the location registration is successful, the selected PLMN becomes the registered PLMN. During registration, if the network declares that the TA is forbidden, the TA is added to a list of forbidden TAs for roaming and is stored in the UE. The UE must then search for a suitable cell in the same PLMN and in a TA that is not in the forbidden lists. If a PLMN is not allowed, the PLMN is added to the forbidden PLMN list that is stored on the USIM and is not deleted, even when the UE is switched off. If the UE cannot find another PLMN, it enters a limited-service state and can only originate emergency calls or receive an ETWS message.

Samsung Ex. 1005, Page 310 of 1147

Figure 5.19 illustrates an inter-frequency (intra-LTE/inter-RAT) cell reselection procedure. The parameters used in the figure are defined as follows [2]:

- $T_{\text{ReselectionRAT}}$: This specifies the cell reselection timer value. For each target E-UTRA frequency and for each RAT (other than E-UTRA), a specific value for the cell reselection timer is defined, which is applicable when evaluating reselection within E-UTRAN or other RAT. This parameter is not sent in *System Information*, rather it is used in reselection rules by the UE for each RAT.
- $\text{Threshold}_{(x)\text{high}}$: This parameter specifies the S_{rxlev} threshold (in dB) used by the UE when reselecting a higher priority RAT/frequency than the current serving frequency.
- $\text{Threshold}_{(x)\text{low}}$: This parameter specifies the S_{rxlev} threshold (in dB) used by the UE when reselecting a lower priority RAT/frequency than the current serving frequency.
- $\text{Threshold}_{(\text{Serving})\text{low}}$: This parameter specifies the S_{qual} threshold (in dB) used by the UE on the serving cell when reselecting a lower priority RAT/ frequency.

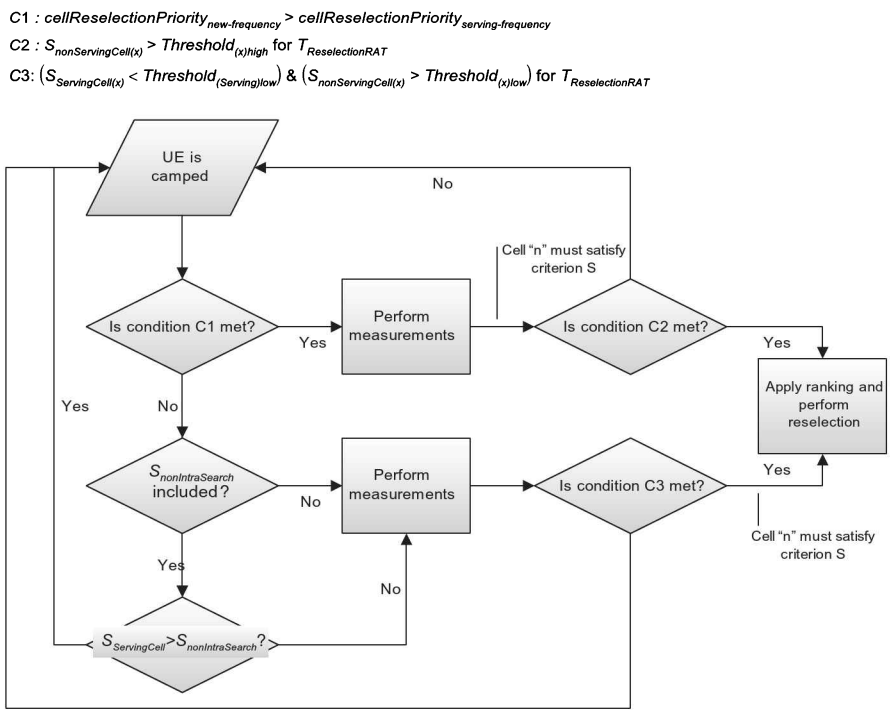


FIGURE 5.19

Illustration of inter-frequency (intra-LTE/inter-RAT) cell reselection procedure (*Note that ranking is only applied to cells with the same reselection priorities*) [1].

- $S_{\text{IntraSearch}}$: This parameter specifies the S_{rxlev} threshold (in dB) for intra-frequency measurements.
- $S_{\text{nonIntraSearch}}$: This parameter specifies the S_{rxlev} threshold (in dB) for E-UTRAN inter-frequency and inter-RAT measurements.

5.5.5 Radio link failure

A UE may declare an RLF in a number of scenarios including the following [1]:

- Downlink physical layer failure when the BLER of PDCCH is greater than 10%: Physical layer problems are detected by monitoring the cell-specific reference signals and estimating the theoretical PDCCH BLER for specific configurations. In-sync and out-of-sync indicators are generated based on the DRX cycle (if configured) and the thresholds Q_{out} (10%) and Q_{in} (2%). If the UE has not been configured with a DRX mode, and the downlink radio link quality estimated during the past 200 ms time interval becomes worse than threshold Q_{out}, the physical layer of the UE sends an out-of-sync indication to the higher layers within 200 ms. When the downlink radio link quality estimated over the past 100 ms time interval becomes better than the threshold Q_{in}, the physical layer of the UE sends an in-sync indication to the higher layers within 100 ms. Two successive indications from the physical layer will be separated by at least 10 ms. The transmitter is turned off within 40 ms after timer T310 expires. If the UE is configured with DRX, the evaluation period for Q_{out} and Q_{in} depends on the DRX configuration. In this case, two successive in-sync or out-of-sync indications are separated by at least the maximum of 10 ms or the DRX cycle.
- Random access problems: If the uplink timing alignment timer expires, then the UE must perform a random access procedure in order to adjust its timing and to obtain an uplink grant before any new data can be transmitted in the uplink. An RLF will be declared, if the UE transmits a maximum number of preamble retransmissions without receiving any response from the eNB.
- Maximum RLC retransmissions: An RLF is declared following the completion of the maximum number of ARQ retransmissions by the RLC sublayer. Following an unsuccessful initial transmission of an RLC PDU, the UE's RLC entity retransmits the PDU a number of times and then declares an RLF upon failure of the receiving entity to confirm reception.
- Handover failure (T304 expiration): Timer T304 is started by the UE after receiving the *RRC Connection Reconfiguration* command ordering a handover. The UE detaches from the source cell and, following successful acquisition of the target cell, uses a random access procedure to gain access to the target cell. If the UE is unable to successfully access the target cell before the expiration of timer T304, an RLF is declared.

Once an RLF has been declared, timer T311 is started. This is only possible, if the AS security was activated prior to the RLF. In this case, SRB1 is suspended and SRB2 and all DRBs are released. If the AS security was not activated, the

UE transitions directly to the idle mode. Upon start of timer T311, the UE attempts cell reselection. If it is unsuccessful before the expiration of timer T311, the UE transitions to the idle mode. If it is successful, the UE starts timer T301 and begins the connection reestablishment procedure with the selected cell. If timer T301 expires without successful reestablishment of the connection, the UE transitions to the idle mode. If successful, SRB1 resumes and the recovery mechanism is complete. The SRB2 and any DRB can then be established.

5.6 RRC timers and constants

A number of LTE/LTE-Advanced RRC functions rely on specific timers to initiate and terminate certain procedures as a result of success or failure of certain tasks, and to measure the run time of certain processes during the operation of the system at the UE and the eNB. The use of timers would prevent a process taking longer than normally expected to perform a task or a series of tasks. The timer would sometimes allow repetition of the associated process/task a permissible number of times before terminating the task. Once a timer associated with a process/task expires when the process has not been successfully concluded, the expiration of the timer terminates the process/task and triggers another process/task. Table 5.7 summarizes the RRC timers specified in LTE Rel-11 specification [1], the conditions to trigger their start and stop, and the actions that the entity would take upon expiration of the timer. Some processes may have several timers associated with them to control different aspects of the process.

5.7 RRC message structure

The RRC messages generally consist of three parts (Figure 5.20):

1. Message type: This information helps the receiver to identify the type of message and other information elements that might be included in the message.
2. RRC transaction identifier: This numerical information is used to separate and identify several ongoing signaling procedures.
3. Other information elements: The RRC message may also include other information elements such as establishment cause, radio resource configuration for SRB1, UE identity, selected PLMN identity, and registered MME information to identify the MME to which the UE is registered.

As an example, the following is an RRC information element that configures the number of transmit antennas and the transmission mode [1,27]:

```
||+-antennaInfo:: = CHOICE [explicitValue] OPTIONAL:Exist
|||  +-explicitValue:: = SEQUENCE [1]
|||     +-transmissionMode:: = ENUMERATED [tm4]
|||     +-codebookSubsetRestriction:: = CHOICE [n2TxAntenna-tm4] OPTIONAL:Exist
|||     |  +-n2TxAntenna-tm4:: = BIT STRING SIZE(6) [111111]
|||     +-ue-TransmitAntennaSelection:: = CHOICE [setup]
|||        +-setup:: = ENUMERATED [closedLoop]
```

Table 5.7 RRC Timers [1]

Timer	Start Condition	Stop Condition	Actions upon Expiration
T300	Transmission of *RRC Connection Request*	Reception of *RRC Connection Setup* or *RRC Connection Reject* message, cell reselection and upon termination of connection establishment by upper layers	Reset MAC, release MAC configuration, and reestablish RLC for all radio bearers that are established and inform upper layers about the failure to establish the RRC connection
T301	Transmission of *RRC Connection Reestablishment Request*	Reception of *RRC Connection Reestablishment* or *RRC Connection Reestablishment Reject* message as well as when the selected cell becomes unsuitable	Go to RRC_IDLE
T302	Reception of *RRC Connection Reject* while performing RRC connection establishment	Upon entering RRC_CONNECTED and upon cell reselection	Inform upper layers about barring mitigation
T303	Access barred while performing RRC connection establishment for mobile originating calls	Upon entering RRC_CONNECTED and upon cell reselection	Inform upper layers about barring mitigation
T304	Reception of *RRC Connection Reconfiguration* message including the *Mobility Control Info* or reception of *Mobility from E-UTRA Command* message including *Cell Change Order*	Criterion for successful completion of handover to E-UTRA or CCO is met (*this criterion is specified in the target RAT in the case of inter-RAT*)	In the case of CCO from E-UTRA or intra E-UTRA handover, initiate the *RRC Connection Reestablishment* procedure
T305	Access barred while performing RRC connection establishment for mobile originating signaling	Upon entering RRC_CONNECTED and upon cell reselection	Inform upper layers about barring mitigation
T310	Upon detecting physical layer problems, i.e., upon	Upon receiving N311 (i.e., *the maximum number of consecutive*	If security is not activated go to RRC_IDLE, else initiate the

(*Continued*)

Table 5.7 (Continued)

Timer	Start Condition	Stop Condition	Actions upon Expiration
	receiving N310 (i.e., *the maximum number of consecutive "out-of-sync" indications received from lower layers*)	*"in-sync" indications received from lower layers*) consecutive in-sync indications from lower layers; upon triggering the handover procedure, and upon initiating the connection reestablishment procedure	connection reestablishment procedure
T311	Upon initiating the *RRC Connection Reestablishment* procedure	Selection of a suitable E-UTRA cell or a cell using another RAT	Enter RRC_IDLE
T320	Upon receiving t320 or upon cell selection/ reselection to E-UTRA from another RAT with validity time configured for dedicated priorities	Upon entering RRC_CONNECTED, when PLMN selection is performed upon request by NAS, or upon cell selection/reselection to another RAT	Discard the cell reselection priority information provided by dedicated signaling
T321	Upon receiving *measConfig* including a *reportConfig* with the purpose set to *reportCGI*. CGI: Cell Global Identification	Upon acquiring the information needed to set all fields of *cellGlobalId* for the requested cell, upon receiving *measConfig* that includes removal of the *reportConfig* with the purpose set to *reportCGI*	Initiate the measurement reporting procedure; stop performing the related measurements and remove the corresponding *measId*
T325	Timer started/restarted upon receiving *RRC Connection Reject* message with *Reprioritization Timer*	–	Stop reprioritization of all frequencies or E-UTRA signaled by *RRC Connection Reject*
T330	Upon receiving *Logged Measurement Configuration* message (*Related to Minimization of Drive Tests feature*)	Upon log volume exceeding the suitable UE memory, upon initiating the release of *Logged Measurement Configuration* procedure	Perform the actions specified in 3GPP TS 36.331
T340	Upon transmitting *UE Assistance Information* message with *powerPrefIndication* set to *normal*	–	–

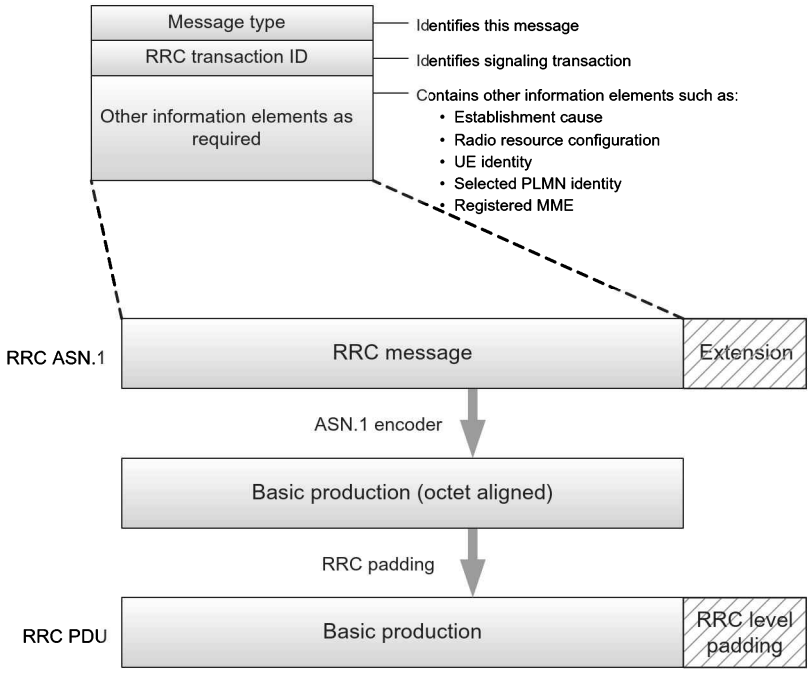


FIGURE 5.20

RRC message structure and processing of RRC messages [1].

The LTE RRC PDU contents are encoded using Abstract Syntax Notation One[3] (ASN.1) as specified in ITU-T Rec. X.680 [23] and X.681 [24]. Transfer syntax for RRC PDUs is derived from their ASN.1 definitions using

[3]ASN.1 is a standard for describing a message that can be sent or received in a network. ASN.1 is divided into two parts: (1) the rules of syntax for describing the contents of a message in terms of data type and content sequence or structure; and (2) how each data item is encoded in a message. ASN.1 is defined in two ISO standards for applications intended for the open systems interconnection framework [23,24]. The following is an example of a message definition specified with ASN.1 notation:

Report:: = SEQUENCE {author OCTET STRING, title OCTET STRING, body OCTET STRING, biblio Bibliography}

In this example, Report is the name of this type of message. SEQUENCE indicates that the message is a sequence of data items. The first four data items have the data type of OCTET STRING, meaning each is a string of bytes. The bibliography data item is another definition named Bibliography that is used.

Bibliography:: = SEQUENCE {author OCTET STRING title OCTET STRING publisher OCTET STRING year OCTET STRING}

Other data types that can be specified include INTEGER, BOOLEAN, REAL, and BIT STRING. An ENUMERATED data type is one that takes one of several possible values. Other data items can be specified as optional.

packed encoding rules as specified in ITU-T Rec. X.691 [25] as shown in Figure 5.20. The "basic production" is obtained by applying unaligned packed encoding rules to the abstract syntax value (the ASN.1 description) as specified in X.691. It always contains a multiple of 8 bits. A transmitter compliant with LTE Rel-11 sets the extension part empty, unless explicitly indicated on a PDU type basis. Transmitters compliant with a later version may send non-empty extensions. If the encoded RRC message does not fill a transport block, the RRC sublayer adds padding bits. This applies to PCCH and BCCH. Padding bits are set to zero and the number of padding bits is a multiple of 8 [1].

References

3GPP Specifications[4]

[1] 3GPP TS 36.331 Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 11), December 2012.

[2] 3GPP TS 36.304, Evolved Universal Terrestrial Radio Access (E-UTRA); UE Procedures in Idle Mode, (Release 11), December 2012.

[3] 3GPP TS 33.401, 3GPP System Architecture Evolution (SAE): Security architecture (Release 12), December 2012.

[4] 3GPP TS 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Overall description, Stage 2 (Release 11), December 2012.

[5] 3GPP TS 36.306, Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Access Capabilities (Release 11), December 2012.

[6] 3GPP TS 36.321, Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol specification (Release 11), December 2012.

[7] 3GPP TS 23.401, General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (Release 11), December 2012.

[8] 3GPP TS 36.101, Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception (Release 11), December 2012.

[9] 3GPP TS 36.104, Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) radio transmission and reception (Release 11), December 2012.

[10] 3GPP TS 36.214, Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer – Measurements (Release 11), December 2012.

[11] 3GPP TS 23.203, Policy and charging control architecture (Release 11), December 2012.

[12] 3GPP TR 36.912, Further Advancements for E-UTRA (LTE-Advanced) (Release 11), September 2012.

[4]3GPP specifications can be found at the following URL: http://www.3gpp.org/ftp/Specs/archive/.

3GPP2 Specifications[5]

[13] 3GPP2 C.S0024-A v3.0, cdma2000 High Rate Packet Data Air Interface Specification, September 2006.

[14] 3GPP2 C.S0057-E v1.0, Band Class Specification for cdma2000 Spread Spectrum Systems, October 2010.

[15] 3GPP2 C.S0005-A v6.0, Upper Layer (Layer 3) Signaling Standard for cdma2000 Spread Spectrum Systems, February 2002.

[16] 3GPP2 A.S0008-C v4.0, Interoperability Specification (IOS) for High Rate Packet Data (HRPD) Radio Access Network Interfaces with Session Control in the Access Network, April 2011.

[17] 3GPP2 C.S0004-A v6.0, Signaling Link Access Control (LAC) Standard for cdma2000 Spread Spectrum Systems, February 2002.

Books

[18] C. Johnson, Long Term Evolution in Bullets, second ed., Create Space Independent Publishing Platform, 2012.

[19] S. Sesia, I. Toufik, M. Baker, LTE, the UMTS Long Term Evolution, From Theory to Practice, second ed., John Wiley & Sons, 2011.

[20] E. Dahlman, S. Parkvall, J. Skold, 4G, LTE/LTE-Advanced for Mobile Broadband, Academic Press, 2011.

[21] H. Holma, A. Toskala, LTE Advanced, 3GPP Solution for IMT-Advanced, John Wiley & Sons, 2012.

[22] H. Holma, A. Toskala, LTE for UMTS, Evolution to LTE-Advanced, John Wiley & Sons, 2011.

Articles

[23] ISO 8824/ITU X.680 specifies the message syntax. <http://www.itu.int/ITU-T/studygroups/com17/languages/X.680-0207.pdf>.

[24] ISO 8825/ITU X.690specifies the basic encoding rules for ASN.1. <http://www.itu.int/ITU-T/studygroups/com17/languages/X.690-0207.pdf>.

[25] ITU-T Recommendation X.691, ASN.1 Encoding Rules: Specification of Packed Encoding Rules (PER), 1999.

[26] LTE Identifiers, NMC Consulting Group, 2011. <http://www.nmcgroups.com>.

[27] LTE System Description. <http://www.sharetechnote.com/>.

[5]3GPP2 specifications can be found at the following URL: http://www.3gpp2.org/Public_html/specs/index.cfm.

This page intentionally left blank

CHAPTER

Packet Data Convergence Protocol Functions

6

CHAPTER CONTENTS

6.1 Ciphering and integrity protection functions 294
6.2 Header compression 296
6.3 Robust header compression 299
6.4 PDCP PDU packet formats 303
6.5 LTE security aspects 304
References 308

The packet data convergence protocol (PDCP) sublayer is part of the LTE layer 2 protocols (see Figure 6.1), which is responsible for the IP header compression of user-plane data packets in order to reduce the number of information bits transmitted over the air-interface. The header compression mechanism is based on the Internet Engineering Task Force (IETF) standard robust header compression (ROHC) [8–11]. The PDCP sublayer is also responsible for ciphering and integrity protection of control-plane RRC messages, as well as in-sequence delivery and duplicate removal. At the receiver side, the PDCP performs the corresponding deciphering and decompression operations. There is one PDCP entity per radio bearer (RB) configured for a terminal. More specifically, the PDCP sublayer provides the following services to other protocol layers on the user-plane: header compression and decompression using ROHC protocol; transfer of user data; in-sequence delivery of upper-layer PDUs at PDCP reestablishment procedure for RLC acknowledged mode (AM); duplicate detection of lower-layer service data units (SDUs) at PDCP reestablishment procedure for RLC AM; retransmission of PDCP SDUs during handover for RLC AM; ciphering and deciphering; and timer-based SDU discarding in the uplink. The main services and functions of the PDCP on the control-plane include ciphering and integrity protection, as well as transfer of control-plane data. Figure 6.1 illustrates the location of PDCP sublayer in the LTE/LTE-Advanced protocol stack.

The LTE radio-access network provides one or more radio bearers to which IP packets are mapped according to their QoS requirements. Note that there are two

LTE-Advanced.

© 2014 Elsevier Inc. All rights reserved.

Samsung Ex. 1005, Page 320 of 1147

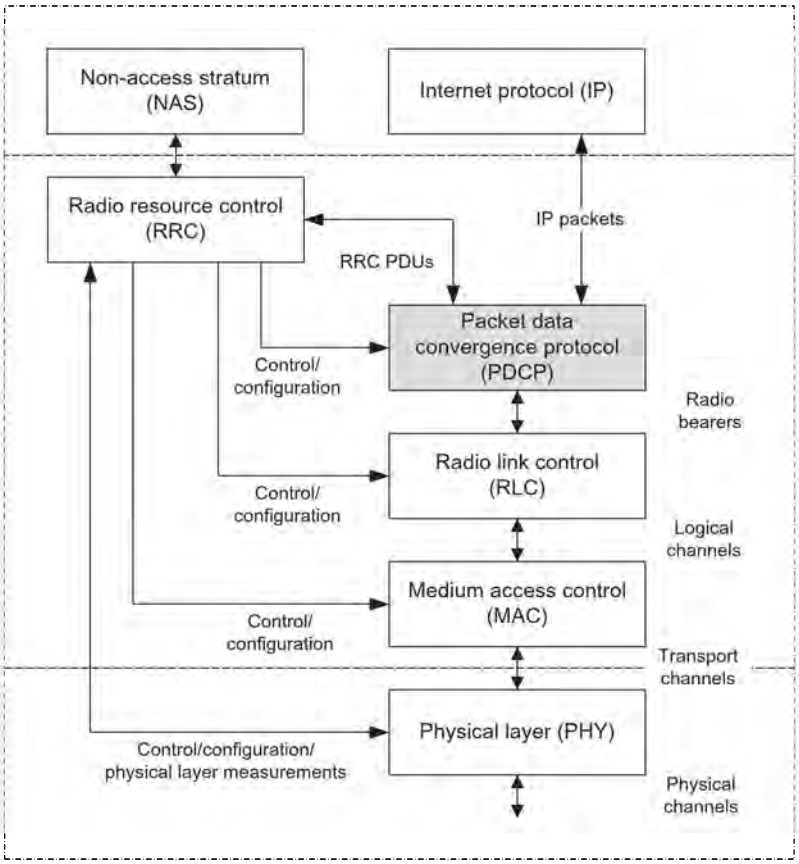


FIGURE 6.1

PDCP sublayer in LTE protocol stack.

types of radio bearers: data radio bearers (DRBs) carrying user-plane data; and signaling radio bearers (SRBs) carrying control-plane information. Each radio bearer is associated with one PDCP entity. Each PDCP entity is associated with one or two (one for each direction) RLC entities depending on the radio bearer characteristic (i.e., unidirectional or bi-directional) and the RLC mode [1–3]. The PDCP entities are located in the PDCP sublayer and are configured by upper layers [2]. A control service access point (SAP) provides the PDCP interface with the RRC sublayer.

As shown in Figure 6.2, the PDCP entities are located in the PDCP sublayer. Several PDCP entities may be defined for a UE. Each PDCP entity carries user-plane data and may be configured to use header compression. Each PDCP entity transports the data of one radio bearer. The LTE standard supports ROHC protocol for header compression as specified by IETF standards [9–11]. Each PDCP entity uses one ROHC instance. A PDCP entity is associated with either the control-plane

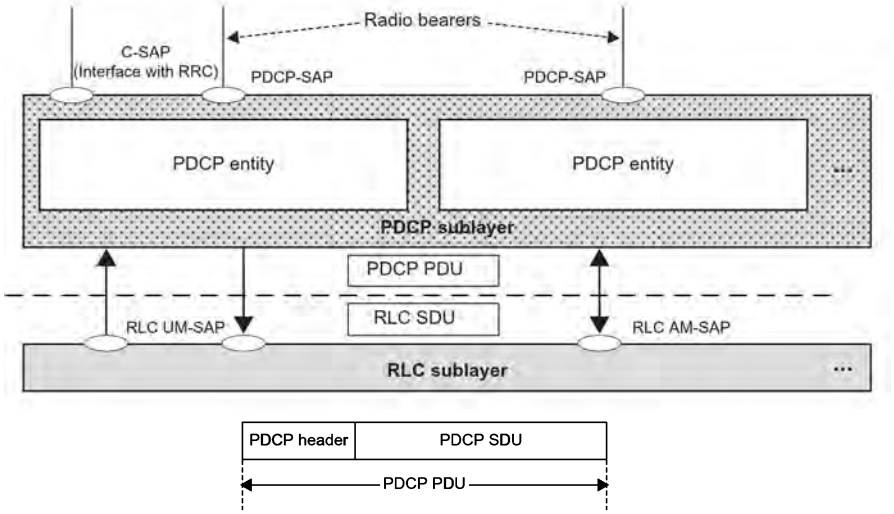


FIGURE 6.2

PDCP sublayer and structure of PDCP PDU [2].

or the user-plane depending on which radio bearer is carrying the data. Figure 6.3 shows the functional decomposition of the PDCP entity in the PDCP sublayer. The figure is based on the radio interface protocol architecture defined in [1]. The PDCP provides services to the RRC sublayer and upper layers in the user-plane at the UE or eNB. As mentioned earlier, PDCP services provided to upper layers include transfer of user-plane and control-plane data, payload header compression, ciphering, and integrity protection.

To provide a better understanding of the PDCP services concerning RRC messages and user data processing, let us further decompose the PDCP functions on the control-plane and user-plane. Figure 6.4 illustrates the PDCP functions of each PDCP entity separately for control-plane and user-plane information. Note that integrity protection and verification functions are only applied to the RRC messages in the downlink and uplink directions, respectively. The header compression and decompression functions are only applied to user-plane data in the downlink and uplink directions, respectively.

The maximum size of a PDCP SDU is 8188 octets [2]. The PDCP uses the services provided by the RLC sublayer. The lower protocol layers provide the following services to the PDCP sublayer:

- Acknowledged data transfer service including indication of successful delivery of PDCP PDUs.
- Unacknowledged data transfer service.
- In-sequence delivery, except at reestablishment of lower layers.
- Duplicate discarding, except at reestablishment of lower layers.

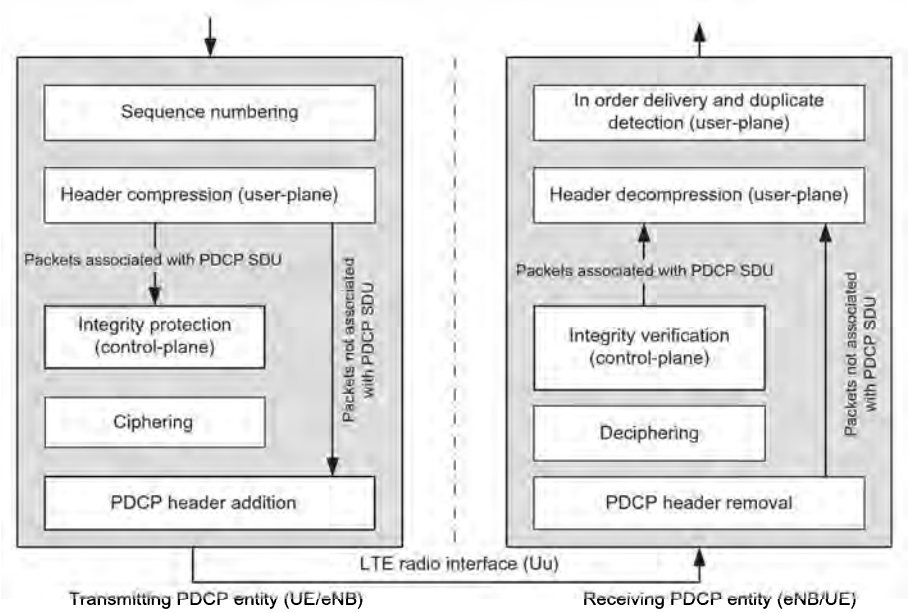


FIGURE 6.3

Functional decomposition of PDCP entities [2].

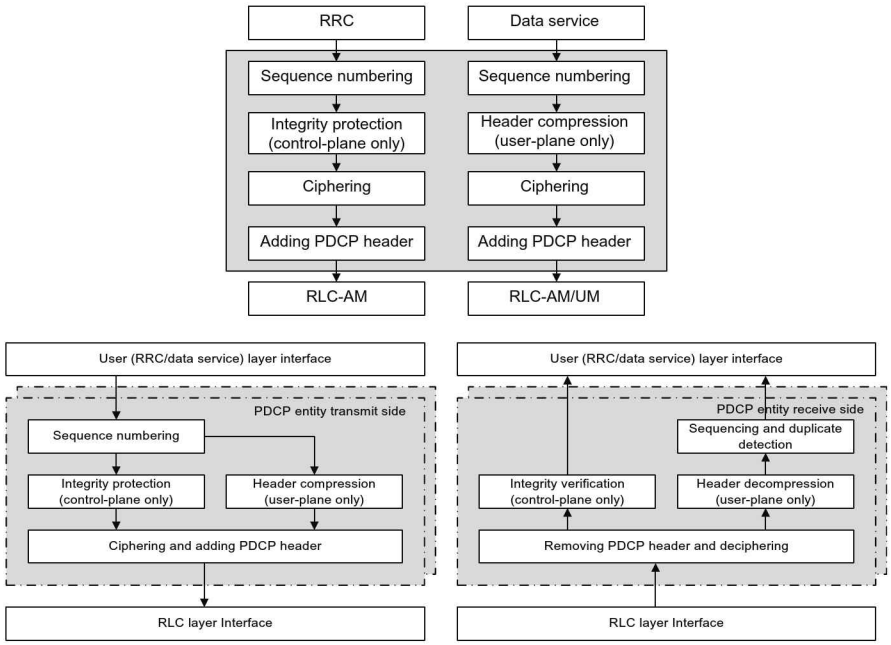


FIGURE 6.4

Illustration of PDCP functions on the control-plane and user-plane.

The PDCP is used for transmission of SRBs (signaling radio bearers carrying control-plane data) and DRBs (data radio bearers carrying user-plane data) that are mapped to dedicated control channel (DCCH) and dedicated traffic channel (DTCH) logical channels. The PDCP is not used for other types of logical channels.

The IETF ROHC protocol is utilized for header compression in LTE [9–11]. There are multiple header compression algorithms, called profiles, defined for the ROHC protocol. Each profile is specific to the particular network layer, transport layer, or upper layer protocol combination, e.g., transmission control protocol/IP (TCP/IP) or real-time transport protocol/user datagram protocol/IP (RTP/UDP/IP) [13,14]. The multiplexing of different flows (with or without header compression) over the ROHC channels, as well as association of a specific IP flow with a specific context state during initialization of the compression algorithm for that flow, are described in [9–11]. The implementation of ROHC functionality and the supported header compression profiles are not specified in 3GPP LTE specifications. Table 6.1 provides the description of ROHC profiles that are used in LTE.

PDCP entities associated with DRBs can be configured by upper layers to use header compression. IETF RFC 4995 provides configuration parameters that are mandatory and must be configured by upper layers between compressor and decompressor entities at the transmitter and receiver sides [11]. Those parameters define the ROHC channel. The ROHC channel is a unidirectional channel, i.e., there is one

Table 6.1 Supported Header Compression Protocols and Profiles [2]

Profile Identifier	Usage	Reference
0×0000	No compression	RFC 4995
0×0001	RTP/UDP/IP	RFC 3095, RFC 4815
0×0002	UDP/IP	RFC 3095, RFC 4815
0×0003	ESP[a]/IP	RFC 3095, RFC 4815
0×0004	IP	RFC 3843, RFC 4815
0×0006	TCP/IP	RFC 4996
0×0101	RTP/UDP/IP	RFC 5225
0×0102	UDP/IP	RFC 5225
0×0103	ESP/IP	RFC 5225
0×0104	IP	RFC 5225

[a]*Encapsulating security payload (ESP) is a key component of the IPsec protocol suite. The ESP protocol provides confidentiality and integrity by encrypting data to be protected and places the encrypted data in the data portion of the IP ESP payload. Depending on the user's security requirements, this mechanism may be used to encrypt either a transport-layer segment (e.g., TCP and UDP), or the entire IP datagram. Encapsulating the protected data is necessary to provide confidentiality for the entire original datagram. ESP also supports encryption-only and authentication-only configurations, but using encryption without authentication is not recommended because of insecurity. Unlike the authentication header (AH), ESP does not protect the IP packet header. However, in the tunnel mode, where the entire original IP packet is encapsulated with a new packet header, ESP protection is applied to the entire inner IP packet (including the inner header), while the outer header remains unprotected. ESP operates directly on top of IP using IP number 50 [15].*

Samsung Ex. 1005, Page 324 of 1147

channel for the downlink and one for the uplink. Therefore, there is one set of parameters for each channel and the same values are used for both channels belonging to the same PDCP entity.

6.1 Ciphering and integrity protection functions

As shown in Figure 6.5, the ciphering function includes both ciphering and deciphering and is performed in the PDCP sublayer. In LTE, the RLC SDUs are ciphered. For the control-plane, the data unit, that is ciphered, is the data part of the PDCP PDU and the message authentication code for integrity (MAC-I), i.e., a 32-bit field that carries the message authentication code. For the user-plane, the data unit that is ciphered is the data part of the PDCP PDU; ciphering is not applicable to the PDCP control PDUs. The ciphering algorithm and key to be used by the PDCP entity are configured by upper layers, and the ciphering method is applied according to the security architecture of 3GPP system architecture evolution [4]. The ciphering function is activated by upper layers. After security activation, the ciphering function is applied to all PDCP PDUs indicated by upper layers for the downlink and the uplink transmissions. The parameters that are required by the PDCP for ciphering are defined in [2], and are input to the ciphering algorithm. The required inputs to the ciphering function include the *COUNT* (a combination of SN and HFN which is 32 bits) and *DIRECTION* (one-bit direction of the transmission). The parameters required by the PDCP which are provided by upper layers are *BEARER* (defined as the RB identifier which is 5 bits) and *KEY* (the 128-bit ciphering keys for the control-plane and for the user-plane) [2].

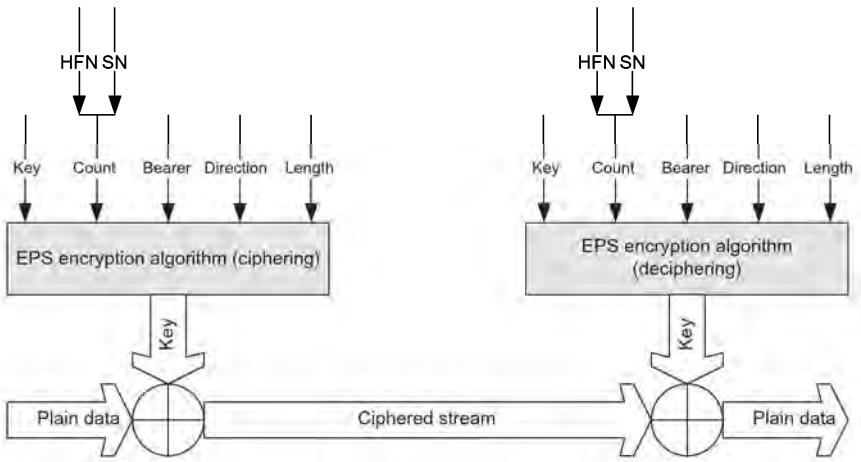


FIGURE 6.5

Illustration of the ciphering and deciphering procedures.

As shown in Figure 6.6, the integrity protection function includes both integrity protection and integrity verification and is performed in the PDCP sublayer for PDCP entities associated with SRBs. The data unit that is integrity protected is the PDU header and the data part of the PDU prior to ciphering. The integrity protection algorithm and key to be used by the PDCP entity are configured by upper layers and the integrity protection method is applied according to the security architecture of 3GPP system architecture evolution [4]. The integrity protection function is activated by upper layers. Following the security activation, the integrity protection function is applied to all PDUs including and subsequent to the PDU indicated by upper layers for the downlink and the uplink transmissions. As the RRC message which activates the integrity protection function is itself integrity protected with the configuration included in that RRC message, the message must be decoded by the RRC sublayer before the integrity protection verification can be performed for the PDU in which the message was received. The parameters that are required by the PDCP for integrity protection are defined in [2], and are input to the integrity protection algorithm. The required inputs to the integrity protection function include the *COUNT* value and *DIRECTION* (direction of the transmission). The parameters required by the PDCP which are provided by upper layers are *BEARER* (defined as the RB identifier) and the *KEY*.

During transmission, the UE computes the value of the MAC-I field, and at reception, it verifies the integrity of the PDCP PDU by calculating the X-MAC, i.e., computed MAC-I, based on the input parameters. If the calculated X-MAC is the same as the received MAC-I, integrity protection is verified successfully. When a PDCP entity receives a PDCP PDU that contains reserved or invalid values, it discards the received PDU. The PDCP data PDU is used to transport the PDCP SDU SN and user-plane data containing an uncompressed PDCP SDU,

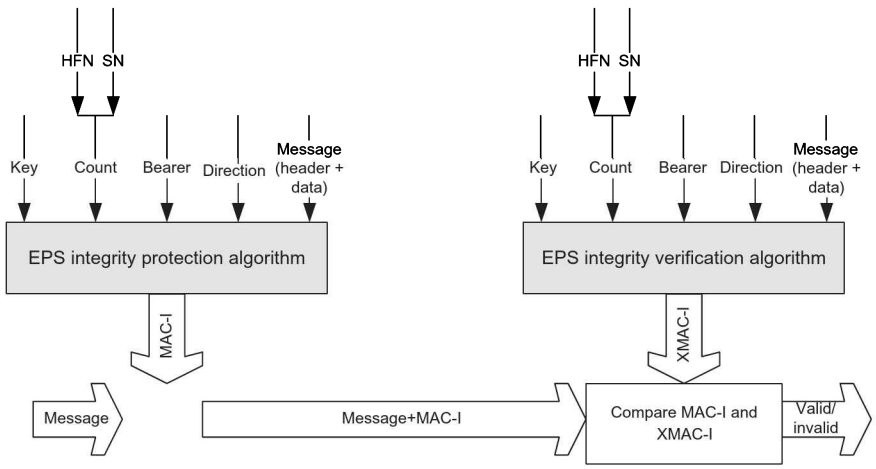


FIGURE 6.6

Integrity protection and verification procedures.

user-plane data containing a compressed PDCP SDU, or control-plane data as well as MAC-I field for SRBs. The PDCP control PDU is used to convey the PDCP status report identifying missing PDCP SDUs following PDCP reestablishment and header compression control information, e.g., ROHC feedback [2].

6.2 Header compression

In some services and applications such as voice over IP (VoIP), interactive gaming, multimedia messaging, etc., the data payload of the IP packet is almost the same size or even smaller than the header itself. Over the end-to-end connection comprising multiple hops, these protocol headers are extremely important but over a single point-to-point link, these headers serve no useful purpose. It is possible to compress these headers, and thus save the bandwidth and use expensive radio resources more efficiently. The header compression also provides other important benefits, such as reduction in packet loss and improved interactive response time [5,8−11]. Payload header compression is the process of suppressing the repetitive portion of payload headers at the sender and restoring them at the receiver side of a low-bandwidth/capacity-limited link. The use of header compression has a well-established history in transport of IP-based payloads over capacity-limited wireless links where more bandwidth efficient transport methods are required. The IETF has developed several header compression protocols that are widely used in telecommunication systems [5,8−11,19−22].

Let us consider some examples of the compression ratios that can be achieved using header compression schemes. The IP version 4 protocol header consists of 20 bytes [6], when combined with the UDP[1] header of 8 bytes and RTP[2] header

[1]The UDP is part of the IP suite, i.e., the set of network protocols used for the Internet. The use of UDP would allow computer applications to send datagrams to other hosts on an IP network without requiring prior signaling to set up data paths. The UDP is defined in IETF RFC 768. The UDP uses a simple transmission model without implicit handshaking dialogues for guaranteeing reliability, ordering, or data integrity. Thus, the UDP provides an unreliable service, and datagrams may be lost or arrive out of order or duplicated. The UDP assumes that error checking and correction is either not necessary or is performed at the application layer, avoiding the overhead of such processing at the network interface level. Delay-sensitive applications such as VoIP and interactive gaming often use UDP because dropping packets would avoid potentially variable and long delays. If error correction schemes are needed at the transport layer, an application may use the TCP which is designed for this purpose. Unlike the TCP, UDP is compatible with packet broadcasting and multicasting. Common applications include media streaming, IPTV, VoIP, and interactive gaming [14,16].

[2]The RTP defines a standardized packet format for delivering audio and video over the Internet. It is defined by IETF RFC 3550 [12]. The RTP is used in communication and entertainment systems that involve media streaming such as telephony, video teleconference applications, and web-based push-to-talk features. Relying on signaling protocols such as session initiation protocol (SIP), the RTP is one of the key ingredients of voice and media transport over IP networks. The RTP is usually used in conjunction with the RTP control protocol (RTCP). While RTP carries the media streams (e.g., audio and video), the RTCP is used to monitor transmission statistics and QoS information. When both protocols are used in conjunction, RTP is usually originated and received on even port numbers, whereas RTCP uses the next highest odd port number [14,17].

of 12 bytes results in an IPv4/UDP/RTP header size of 40 bytes. A header compression scheme typically compresses such headers to 2–4 bytes in the steady-state. Note that the RTP payload size of some commonly used voice codecs is approximately 20–40 bytes for active speech; therefore, the 40 byte IPv4/UDP/RTP protocol header size would be a relatively large overhead. Using header compression in such cases would result in major bandwidth savings. The amount of overhead for small packets when using IP version 6 (IPv6), with increased header size of 40 bytes due to increased IP address space [7], would be even larger. In low-bandwidth or congested networks, the use of header compression may yield better response times due to smaller packet sizes. A small packet may further reduce the probability of packet loss [21]. It has been observed that in applications such as video transmission over wireless links, the use of header compression does not improve the video quality in spite of lower bandwidth usage. For voice transmission, the voice quality may improve while utilizing a lower transmission bandwidth. In summary, header compression helps improve network transmission efficiency, quality, and speed by decreasing protocol header overhead, reducing the packet loss, decreasing interactive response time, and increasing network core users per channel bandwidth which means less infrastructure deployment costs. Figure 6.7 illustrates the structure of IPv6, UDP, and RTP headers.

The IP together with transport protocols such as TCP or UDP and application-layer protocols (e.g., RTP) are described in the form of payload headers. The information carried in the header helps the applications to communicate over large distances connected by multiple links or hops in the network. This information consists of source and destination addresses, ports, protocol identifiers, SNs, error checksums, etc. Under nominal conditions, most of the information carried in packet headers remains the same or changes in specific patterns. By observing the fields that remain constant or change in specific patterns, it is possible either not to send them in each packet, or to represent them in a smaller number of bits than would have been originally required. This process is referred to as compression.

The process of header compression uses the concept of flow context, which is a collection of information about field values and change patterns of field values in the packet header. This context is formed on the compressor and the decompressor side for each packet flow. The first few packets of a newly identified flow are used to build the context on both sides. These packets are sent without compression. The number of these first few packets, which are initially sent uncompressed, is closely related to link characteristics like bit error rate and round trip time. Once the context is established on both sides, the compressor compresses the payload headers as much as possible. By taking into account the link conditions and feedback from the decompressor, the compressed packet header sizes may vary. At certain intervals and in the case of error recovery, uncompressed packet headers are sent to reconstruct the context and revert back to normal operational mode, which is sending compressed packet headers. The header

Samsung Ex. 1005, Page 328 of 1147

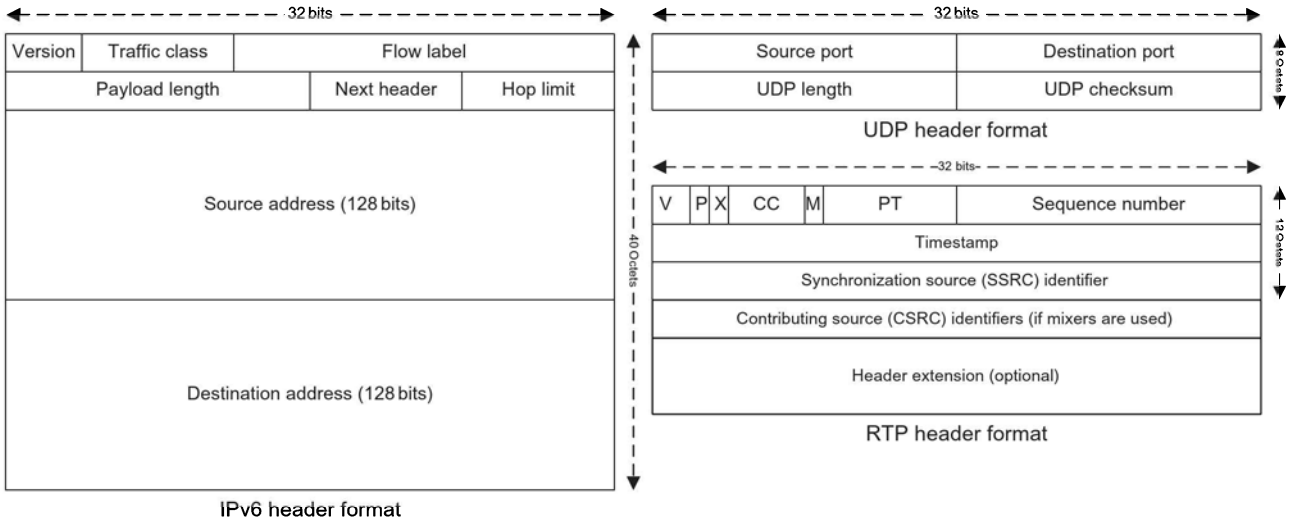


FIGURE 6.7

IPv6, UDP, and RTP header formats [13].

compression module is a part of the protocol stack on the devices. It is a feature which must be negotiated before it can be used on a link. Both end points must agree if they support header compression and on the related parameters to be negotiated.

Considering the end-to-end connectivity over IP, the header compression does not introduce any changes in the data payload when it compresses and decompresses the header. The header compression is a hop-to-hop process and is not applied in an end-to-end connection. At each hop in the IP network, it becomes necessary to decompress the packet to be able to perform operations such as routing, QoS negotiation, and parameter adjustment, etc. Header compression is best suited for specific links in the network characterized by relatively low bandwidth, high bit error rates, and long round trip times. The performance of a header compression scheme can be described with three parameters [18,19]: compression efficiency; robustness; and compression transparency. The compression efficiency is determined by how much the header sizes are reduced by the compression scheme. The compression transparency is a measure of the extent to which the scheme ensures that the decompressed headers are semantically identical to the original headers. If all decompressed headers are semantically identical to the corresponding original headers, the transparency is considered to be 100%. Compression transparency is high when damage propagation is low. When the context of the decompressor is not consistent with the context of the compressor, decompression may fail to reproduce the original header. This condition can occur when the context of the decompressor has not been initialized properly, or when packets have been lost or damaged between compressor and decompressor.

6.3 Robust header compression

The ROHC scheme reduces the size of the transmitted IP/UDP/RTP header by removing redundancies. This mechanism starts by classifying header fields into different classes according to their variation pattern. The fields that are classified as *inferred* are not sent. The *static* fields are sent initially and then are not sent again, and the fields with varying information are always sent. The ROHC mechanism is based on a context[3] which is maintained by both ends, i.e., the compressor and the decompressor (see Figure 6.8). The context encompasses the entire header and ROHC information. Each context has a context identifier, which

[3]The context of the compressor is the state it uses to compress a header. The context of the decompressor is the state it uses to decompress a header. Either of these or the combinations of these two are usually referred to as "context." The context contains relevant information from previous headers in the packet stream, such as static fields and possible reference values for compression and decompression. Moreover, additional information describing the packet stream is also part of the context, for example information about how the IP Identifier field changes and the typical inter-packet increase in sequence numbers or timestamps [14].

Samsung Ex. 1005, Page 330 of 1147

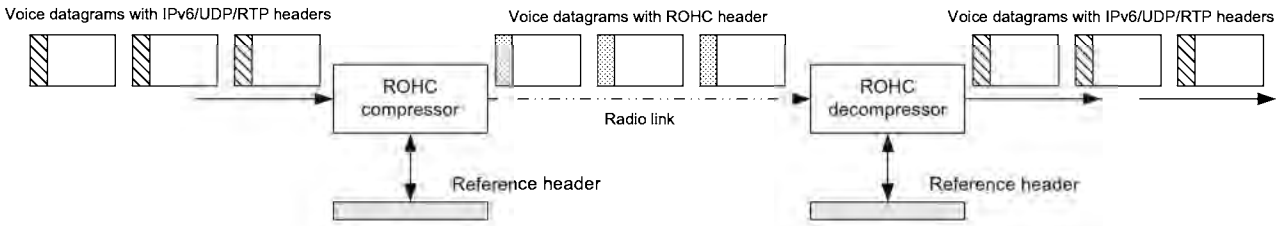


FIGURE 6.8

Example ROHC compression/decompression of IP/UDP/RTP headers for communication over a radio link.

identifies the flows. The ROHC scheme operates in one of the following three operational modes [8–11]:

1. Unidirectional mode (U): In the unidirectional mode of operation, packets are only sent in one direction, from compressor to decompressor. This mode therefore makes ROHC usable over links where a return path from decompressor to compressor is unavailable or undesirable.
2. Optimistic mode (O): The bi-directional optimistic mode is similar to the unidirectional mode, except that a feedback channel is used to send error recovery requests and (optionally) acknowledgments of significant context updates from decompressor to compressor. The O-mode aims to maximize compression efficiency and sparse usage of the feedback channel.
3. Reliable mode (R): The bi-directional reliable mode differs in many ways from the previous two. The most important differences are a more intensive usage of the feedback channel and a stricter logic at both compressor and decompressor that prevents loss of context synchronization between compressor and decompressor except for very high residual bit error rates.

The U-mode is used when the link is unidirectional or when feedback is not possible. For bi-directional links, the O-mode uses positive feedback packets (acknowledgment (ACK)) and R-mode uses positive and negative feedback packets (ACK and negative acknowledgment (NACK)). ROHC always starts header compression using the U-mode, even if it is used in a bi-directional link. ROHC does not make retransmission when an error occurs and the erroneous packet is dropped. The ROHC feedback is used only to indicate to the compressor side that there was an error and probably the context was damaged. After receiving a negative feedback, the compressor always reduces its compression level.

The ROHC compressor has three compression states as follows [8–11]:

1. Initialization and refresh (IR) where the compressor has just been created or reset and full packet headers are sent.
2. First order (FO) where the compressor has detected and stored the static fields such as IP addresses and port numbers on both sides of the connection.
3. Second order (SO) where the compressor is suppressing all dynamic fields such as RTP SNs, and sending only a logical SN and partial checksum to cause the other side to generate the headers based on prediction and to verify the headers of the next expected packet.

Each compression state uses a different header format in order to send the header information. The IR compression state establishes the context, which contains static and dynamic header information. The FO compression state provides the change pattern of dynamic fields. The SO compression state sends encoded values of SN and time stamp (TS), forming the minimal size packets (see Figure 6.9). Using this header format, all header fields can be generated at the other end of the radio link using the previously established change pattern. When some updates or errors occur, the compressor returns to upper compression states.

Samsung Ex. 1005, Page 332 of 1147

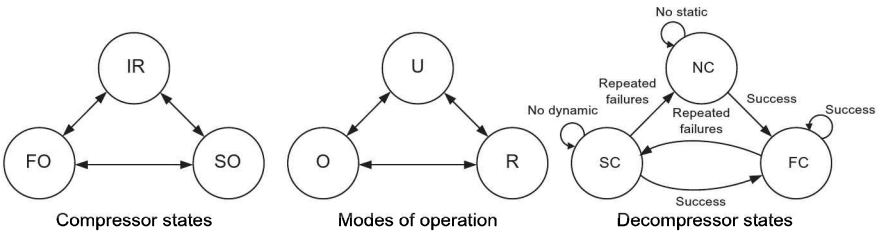


FIGURE 6.9

ROHC state machines [8].

It only transitions to the SO compression state after retransmitting the updated information and reestablishing the change pattern in the decompressor.

In the U-mode, the feedback channel is not used. To increase the compression level, an optimistic approach is used for the compressor to ensure that the context has been correctly established at the decompressor side. This means that the compressor uses the same header format for a number of packets. Since the compressor does not know whether the context is lost, it also uses two timers, to be able to return to the FO and IR compression states. The decompressor works at the receiving end of the link and decompresses the headers based on the header fields' information of the context. Both compressor and decompressor use a context to store all the information about the header fields. To ensure correct decompression, the context should be always synchronized. The decompressor has three states as follows: (1) no context (NC) where there is no context synchronization; (2) static context (SC) where the dynamic information of the context has been lost; and (3) full context (FC) when the decompressor has all the information about header fields. In the FC state, the decompressor transitions to the initial states as soon as it detects corruption of the context. The decompressor uses the "*k out of n*" rule by looking at the last *n* packets with cyclic redundancy check (CRC) failures. If *k* CRC failures have occurred, it assumes the context has been corrupted and transitions to an initial state (SC or NC). The decompressor also sends feedback according to the operation mode (see Figure 6.9).

The values of the ROHC compression parameters that determine the efficiency and robustness are not defined in ROHC specification and are not negotiated initially but are stated as implementation dependent. The values of these parameters stay fixed during the compression process. Those compression parameters are as follows [8–11]:

- *L*: In U-mode and O-mode the ROHC compressor uses a confidence variable *L* in order to ensure the correct transmission of header information.
- *Timer_1* (*IR_TIMEOUT*): In U-mode, the compressor uses this timer to return to the IR compression level and periodically resends static information.
- *Timer_2* (*FO_TIMEOUT*): The compressor also uses another timer in U-mode and this timer is used to go downward to the FO compression level if the compressor is working in the SO compression level.

- *Sliding Window Width* (*SWW*): The compressor while compressing header fields like *Sequence Number* and *Time Stamp* uses Window-based least significant bit (*W_LSB*) encoding that uses a sliding window of width equal to *SWW*.
- *W_LSB* encoding is used to compress those header fields whose change pattern is known. When using this encoding, the compressor sends only the least significant bits. The decompressor uses these bits to construct the original value of the encoding fields.
- *k* and *n*: The ROHC decompressor uses a "*k out of n*" failure rule, where *k* is the number of packets received with an error in the last *n* transmitted packets. This rule is used in the state machine of the decompressor to determine the damage of the context and to move to an initial state after sending a negative acknowledgment to the compressor, if a bi-directional link is used. The decompressor does not declare a context corruption and remains in its current state until *k* packets arrive with error in the last *n* packets.

6.4 PDCP PDU packet formats

The PDCP PDU packets are used to transport user-plane data or control-plane information corresponding to the RRC sublayer or NAS. Depending on the type of information, a PDCP PDU packet may include a PDCP SDU SN and user-plane data containing an uncompressed PDCP SDU; or user-plane data containing a compressed PDCP SDU; or control-plane data and a MAC-I field for SRBs. The PDCP control PDU conveys a PDCP status report indicating which PDCP SDUs are missing following PDCP reestablishment or header compression control information, e.g., interspersed ROHC feedback. Figure 6.10 illustrates various forms of PDCP PDU packets [2].

As shown in Figure 6.10, the PDCP PDU packet is a bit string that is octet-aligned. The bit string is supposed to be read from left to right, and then in the reading order of the lines. The bit order of each parameter field within a PDCP PDU is represented with the first and most significant bit in the leftmost bit, and the last and least significant bit in the rightmost bit. The PDCP SDUs are bit strings that are octet-aligned. A compressed or uncompressed SDU is included in a PDCP PDU from the first bit onward.

The PDCP PDU for carrying control-plane information may contain the signaling messages from the RRC and NAS layers where the signaling messages are carried on either SRB 1 or SRB 2 and there is a 5-bit SN field and 32-bit MAC-I checksum field for integrity protection attached to the end of the PDCP control-plane PDU.

The PDCP PDU for carrying user-plane information may contain DRB. A one-bit D/C field (D: user-plane data; C: control information generated at the PDCP layer) is included in the PDCP PDUs carrying user-plane information.

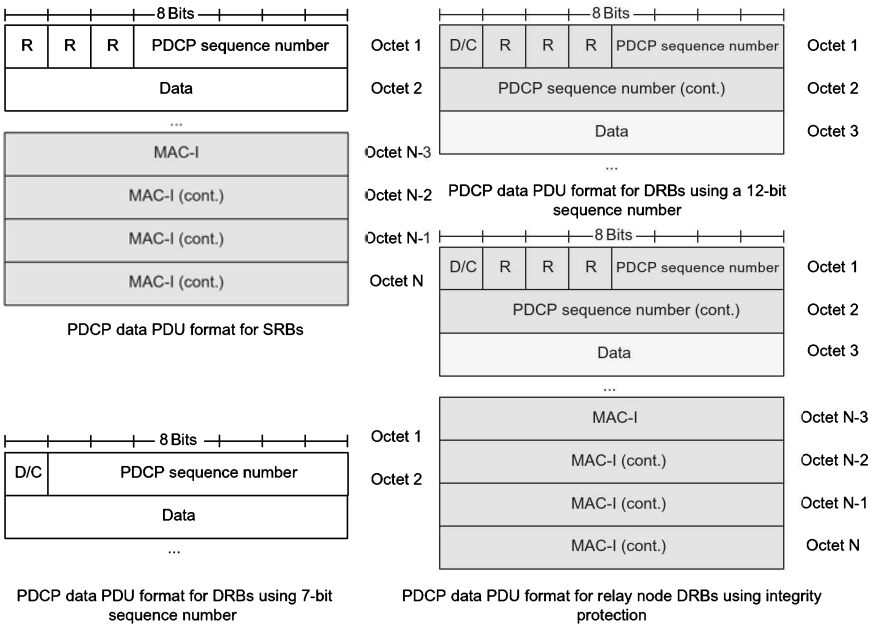


FIGURE 6.10

PDCP PDU packet formats [2].

There are two formats corresponding to a 7-bit or a 12-bit sequence number, and the length of the sequence number is configured on a DRB basis. The unused fields are marked as reserved.

The PDCP data PDU with a 12-bit SN is used for carrying data from DRBs mapped to RLC AM or RLC UM SDUs. The PDCP data PDU with a 7-bit SN is used for transporting data from DRBs mapped on RLC UM. The PDCP control PDU is used to carry one interspersed ROHC feedback packet and is applicable to DRBs mapped to RLC AM or RLC UM SDUs. The PDCP control PDU is used to transport one PDCP status report when a 12-bit SN is utilized. Another form of the PDCP control PDU may be utilized to carry one PDCP status report when a 15-bit SN is utilized, which is applicable to DRBs mapped to RLC AM SDUs. The PDCP data PDU may be used to carry DRBs for relay nodes when integrity protection is configured. The latter format shown in Figure 6.10 is applicable to DRBs mapped to RLC AM or RLC UM SDUs [2].

6.5 LTE security aspects

The E-UTRAN security has been designed based on the following principles. The keys used for NAS and AS protection are dependent on the algorithm with which they are used [1,4]. The eNB keys are cryptographically separated from the EPC

keys that are used for NAS protection, making it impossible to use the eNB key to drive an EPC key. The AS and NAS keys are derived in the EPC/UE from key material that was generated by a NAS (EPC/UE) level authentication and key agreement (AKA) procedure[4] (K_{ASME}), and is identified with a key identifier (KSI_{ASME}). The eNB key (K_{eNB}) is sent from the EPC to the eNB when the UE is entering the ECM-CONNECTED state, i.e., during RRC connection or S1 context setup. Separate AS and NAS level security mode procedures are used. The AS level security mode procedure configures AS security (i.e., the RRC and user-plane), and NAS level security mode procedure configures NAS security. Both integrity protection and ciphering for the RRC are activated within the same AS security mode command procedure, but not necessarily within the same message [1].

The user-plane ciphering is activated at the same time as the RRC ciphering. The keys that are stored in the eNBs never leave a secure environment, and user-plane data ciphering/deciphering is conducted within the secure environment where the related keys are stored. The information related to the eNB keys is sent between the eNBs during ECM-CONNECTED state in intra-E-UTRAN mobility. An SN denoted as *COUNT* is used as input to the ciphering and integrity protection. A given SN must only be used once for a given eNB key, except for identical retransmission on the same radio bearer in the same direction. The same SN can be used for both ciphering and integrity protection. An HFN, i.e., an overflow counter mechanism, is used in the eNB and UE in order to limit the actual number of SN bits that are needed to be sent over the radio air-interface. The HFN is synchronized between the UE and eNB. If corruption of keys is detected, the UE has to restart the radio level attach procedure, e.g., similar radio level procedure to the RRC_IDLE to RRC_CONNECTED mode transition or initial attach. Since subscriber identity module (SIM) access is not supported in E-UTRAN, an idle-mode UE not equipped with USIM cannot attempt to reselect the E-UTRAN. To prevent handover to the E-UTRAN, the UE which is not equipped with USIM indicates E-UTRA support in UE capability signaling in other radio access

[4]The AKA mechanism is based on challenge-response mechanisms and symmetric cryptography. The AKA scheme is typically used in a UMTS universal subscriber identity module (USIM), or a cdma2000 removable user identity module (RUIM). The third generation AKA provides substantially longer key lengths and mutual authentication compared to the second generation mechanisms such as GSM AKA. The introduction of AKA inside EAP further enables several new applications, including the use of the AKA as a secure PPP authentication method in devices that already contain an identity module, the use of the third generation mobile network authentication infrastructure in the context of wireless LANs as well as relying on AKA and the existing infrastructure in a seamless way with any other technology that can use EAP. In AKA, the identity module and the home environment have agreed on a secret key in advance. The "home environment" refers to the home operator's authentication network infrastructure. Furthermore, the actual authentication process starts by having the home environment produce an authentication vector, based on the secret key and a sequence number. The authentication vector contains a random part, an authenticator part used for authenticating the network to the identity module, an expected result part, a 128-bit session key for integrity check, and a 128-bit session key for encryption [1,4].

Samsung Ex. 1005, Page 336 of 1147

technologies (RATs). A simplified key derivation algorithm is illustrated in Figure 6.11, where [1]:

- K_{NASint} is a key, which is only used for the protection of NAS traffic with a particular integrity algorithm. This key is derived by the UE and MME from K_{ASME}, as well as an identifier for the integrity algorithm.
- K_{NASenc} is a key, which is only used for the protection of NAS traffic with a particular encryption algorithm. This key is derived by the UE and MME from K_{ASME}, as well as an identifier for the encryption algorithm.
- K_{eNB} is a key derived by the UE and MME from K_{ASME}. K_{eNB} may also be derived by the target eNB from the next hop (NH) during handover. K_{eNB} is used for the derivation of K_{RRCint}, K_{RRCenc}, and K_{UPenc}, for the derivation of K_{eNB^*} upon handover.
- K_{eNB^*} is a key derived by the UE and source eNB from either K_{eNB}, or from a fresh NH. K_{eNB^*} is used by the UE and target eNB as a new K_{eNB} for RRC and user-plane traffic.
- K_{UPenc} is a key which is only used for the protection of user-plane traffic with a particular encryption algorithm. This key is derived by the UE and eNB from K_{eNB}, as well as an identifier for the encryption algorithm.
- K_{RRCint} is a key which is only used for the protection of RRC traffic with a particular integrity algorithm. K_{RRCint} is derived by the UE and eNB from K_{eNB}, as well as an identifier for the integrity protection algorithm.

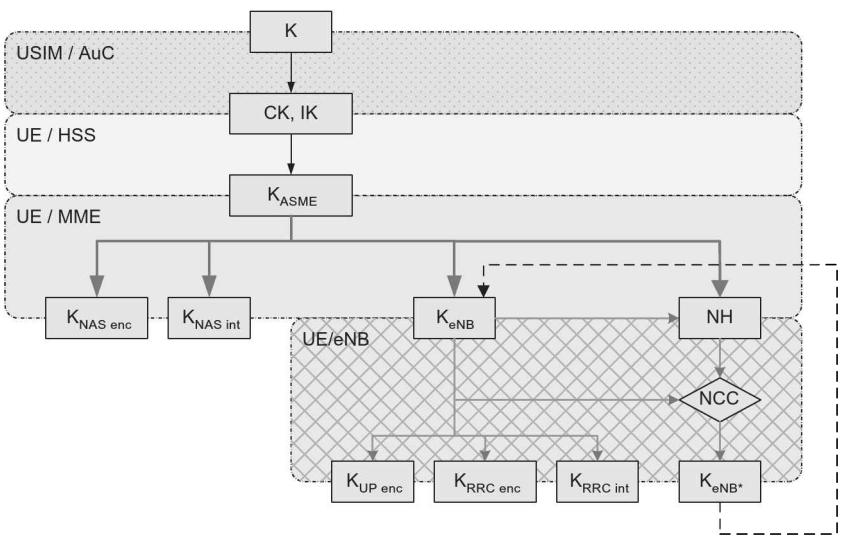


FIGURE 6.11

Key derivation procedure in LTE [1].

- K_{RRCenc} is a key which is only used for the protection of RRC traffic with a particular encryption algorithm. K_{RRCenc} is derived by the UE and eNB from K_{eNB} as well as an identifier for the encryption algorithm.
- The NH is used by the UE and eNB in the derivation of K_{eNB^*} for the provision of *Forward Security*. The NH is derived by the UE and MME from K_{ASME} and K_{eNB} when the security context is established, or from K_{ASME} and the previous NH.
- Next hop chaining count (NCC) is a counter related to NH, i.e., the amount of key chaining that has been performed, which allows 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.

The AuC, HSS, and MME entities shown in Figure 6.11 are authentication center (i.e., the network entity that can authenticate subscribers in a mobile network), home subscriber service (i.e., the main IP multimedia sub-system (IMS) database which also acts as a database in EPC that combines legacy home location register and AuC functions for circuit-switched and packet-switched domains), and mobility management entity, respectively [1].

In Figure 6.11, CK and IK denote keys that have been agreed during the AKA procedure. The MME invokes the AKA procedures by requesting authentication vectors from the home environment (HE), if no unused evolved packet system (EPS) authentication vectors have been stored. The HE sends an authentication response back to the MME that contains a fresh authentication vector including a base-key named K_{ASME}. Consequently, the EPC and the UE share K_{ASME}. From K_{ASME}, the NAS keys, and indirectly K_{eNB} keys and NH 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 state. From the K_{eNB}, the eNB and UE can derive the user-plane and RRC keys.

The RRC and user-plane keys are refreshed during handover. K_{eNB^*} is derived by the UE and source eNB from target *physical cell identifier*, target frequency and K_{eNB}, or alternatively from target *physical cell identifier*, target frequency and NH. K_{eNB^*} is then used as a new K_{eNB} for RRC and user-plane traffic at the target. If the UE transitions to the ECM-IDLE state, all keys are deleted in the eNB. The reuse of COUNT is avoided for the same radio bearer identity in the RRC_CONNECTED mode without K_{eNB} change, and this is left to eNB implementation. If HFN becomes unsynchronized in the RRC_CONNECTED mode between the UE and eNB, the UE is forced to transition to the idle mode. Table 6.2 shows the security termination points in LTE.

Integrity protection for user-plane is not required and thus it is not supported between the UE and serving gateway, or for the transport of user-plane data between the eNB and the serving gateway on the S1 interface (i.e., the interface between eNB and MME). In general, upon RRC_IDLE to RRC_CONNECTED state transition, RRC protection and user-plane protection keys are generated, while keys for NAS protection as well as higher layer keys are assumed to be already available in the MME. These higher layer keys may have been established in the MME as a

Samsung Ex. 1005, Page 338 of 1147

Table 6.2 Security Termination Points in LTE [1]

	Ciphering	Integrity Protection
NAS signaling	Required and terminated in MME	Required and terminated in MME
User-plane data	Required and terminated in eNB	Not required
RRC signaling (AS)	Required and terminated in eNB	Required and terminated in eNB
MAC signaling (AS)	Not required	Not required

result of an AKA run, or transferred from another MME during handover or idle mode mobility. Upon RRC_CONNECTED to RRC_IDLE state transition, the eNB deletes the keys that it stores such that the context for idle-mode UE only resides in the MME. The state and current keys of the UE are assumed to have been deleted at the eNB upon such transition. In particular, the eNB and UE delete NH, K_{eNB}, K_{RRCenc}, K_{RRCint}, and K_{UPenc} and related NCC, while MME and UE keep K_{ASME}, K_{NASint}, and K_{NASenc}.

The RRC and user-plane keys are derived based on the algorithm identifiers and K_{eNB} which results in new RRC and user-plane keys in each handover. Source eNB and UE independently create K_{eNB*}. The K_{eNB*} is provided to target eNB during the handover preparation time. Both target eNB and UE consider the new K_{eNB} to be equal to the received K_{eNB*}. If AS keys (K_{UPenc}, K_{RRCint}, and K_{RRCenc}) need to be changed in RRC_CONNECTED state, an intra-cell handover is used. Inter-RAT handover from UTRAN to E-UTRAN is only supported after activation of integrity protection in UTRAN.

References

3GPP Specifications[5]

[1] 3GPP TS 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 11), December 2012.

[2] 3GPP TS 36.323, Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) specification (Release 11), December 2012.

[3] 3GPP TS 36.322, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification (Release 11), December 2012.

[4] 3GPP TS 33.401, 3GPP System Architecture Evolution (SAE): Security architecture (Release 12), December 2012.

[5]3GPP specifications can be downloaded from http://www.3gpp.org/ftp/Specs/archive/.

IETF Specifications[6]

[5] IETF RFC 2507, M. Degermark et al., IP Header Compression. <http://www.rfc-editor.org/rfc/rfc2507.txt>, February 1999.
[6] IETF RFC 791, Internet Protocol. <http://www.ietf.org/rfc/rfc0791.txt>, September 1981.
[7] IETF RFC 2460, S. Deering, R. Hinden, Internet Protocol, Version 6 (IPv6) specification, December 1998.
[8] IETF RFC 3095, C. Bormann, et al., RObust Header Compression (ROHC): Framework and four profiles: RTP, UDP, ESP, and uncompressed. <http://www.ietf.org/rfc/rfc3095.txt>, July 2001.
[9] IETF RFC 3759, L-E. Jonsson, RObust Header Compression (ROHC): Terminology and Channel Mapping Examples. <http://www.ietf.org/rfc/rfc3759.txt>, April 2004.
[10] IETF RFC 4815, L.-E. Jonsson et al., RObust Header Compression (ROHC): Corrections and Clarifications to RFC 3095, February 2007.
[11] IETF RFC 4995, L.-E. Jonsson et al., The RObust Header Compression (ROHC) Framework, July 2007.
[12] IETF RFC 3550, H. Schulzrinne et al., RTP: A Transport Protocol for Real-Time Applications, July 2003.

Books

[13] D.E. Comer, Internetworking with TCP/IP Vol. 1: Principles, Protocols, and Architecture, fifth ed., Prentice Hall, July 2005.
[14] C. Perkins, RTP: Audio and Video for the Internet, Addison-Wesley Professional, June 2003.

Articles

[15] Encapsulating security payload (ESP). <http://en.wikipedia.org/wiki/IPsec#Encapsulating_Security_Payload>.
[16] User datagram protocol (UDP), Wikipedia. <http://en.wikipedia.org/wiki/User_Datagram_Protocol>.
[17] Real-time transport protocol (RTP), Wikipedia. <http://en.wikipedia.org/wiki/Real-time_Transport_Protocol>.
[18] A. Minaburo, et al., Proposed behavior for robust header compression over a radio link, IEEE Int. Conf. Commun. 7 (2004).
[19] D.E. Taylor, et al., Robust header compression (ROHC) in next-generation network processors, IEEE/ACM Trans. Network. 13 (4) (2005).
[20] S. Ayed, et al., Enhancing robust header compression over IEEE 802 networks, IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, 2006. (WiMob'2006), June 2006.
[21] A. Minaburo, L. Nuaymi, et al., Configuration and analysis of robust header compression in UMTS, 14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications (PIMRC-2003), vol. 3, September 2003.
[22] P. Fortuna, M. Ricardo, Header compressed VoIP in IEEE 802.11, IEEE Wireless Commun. 16 (2009).

[6]IETF specifications can be downloaded from http://www.ietf.org/rfc.html.

Samsung Ex. 1005, Page 340 of 1147

This page intentionally left blank

CHAPTER

Radio Link Control Functions 7

CHAPTER CONTENTS

7.1 RLC architecture 314
7.2 RLC transfer modes 316
7.2.1 Transparent mode 317
7.2.2 Unacknowledged mode 319
7.2.3 Acknowledged mode 321
7.3 RLC ARQ procedure 327
7.3.1 ARQ principles 327
7.3.2 ARQ operation 330
7.4 RLC PDU packet formats 332
7.4.1 Transparent mode PDU format 332
7.4.2 Unacknowledged mode PDU formats 334
7.4.3 Acknowledged mode PDU formats 335
7.4.4 Status PDU format 336
7.5 RLC variables, constants, and timers 337
References 340

The radio link control (RLC) sublayer is located between packet data convergence protocol (PDCP) and medium access control (MAC) sublayers as shown in Figure 7.1. The RLC sublayer is responsible for transferring upper layer protocol data units (PDUs) generated by radio resource control (RRC) or PDCP sublayers [3,4]. The RLC sublayer can operate in one of the three modes of operation defined as transparent mode (TM), unacknowledged mode (UM), and acknowledged mode (AM). Depending on the mode of operation, the RLC entity controls the usage of error correction, concatenation, segmentation, resegmentation, duplicate detection, and in-sequence delivery of service data units. Other functions of the RLC sublayer include protocol error detection, recovery, and service data unit (SDU) discarding. The RLC entity can be reestablished by the RRC such that a reset will cause the transmit/receive buffers to discard any non-transmitted or

LTE-Advanced.

© 2014 Elsevier Inc. All rights reserved.

Samsung Ex. 1005, Page 342 of 1147

incompletely transmitted SDUs. Furthermore, RLC can also be instructed to discard any unmapped RLC SDU by the upper layers such as PDCP [10].

The RLC sublayer communicates with the PDCP sublayer through a service access point (SAP), and with the MAC sublayer via logical channels [1]. The RLC sublayer reformats PDCP PDUs in order to fit them into the payload size permissible by the MAC sublayer, i.e., the RLC transmitter segments and/or concatenates the PDCP PDUs, and the RLC receiver reassembles the RLC PDUs to reconstruct the PDCP PDUs. In addition, the RLC reorders the RLC PDUs if they are received out of sequence due to a hybrid automatic repeat request (HARQ) operation performed in the MAC sublayer. The functions of the RLC sublayer are performed by RLC entities. An RLC entity is configured in one of three data transmission modes: transparent mode, unacknowledged mode, or acknowledged mode.

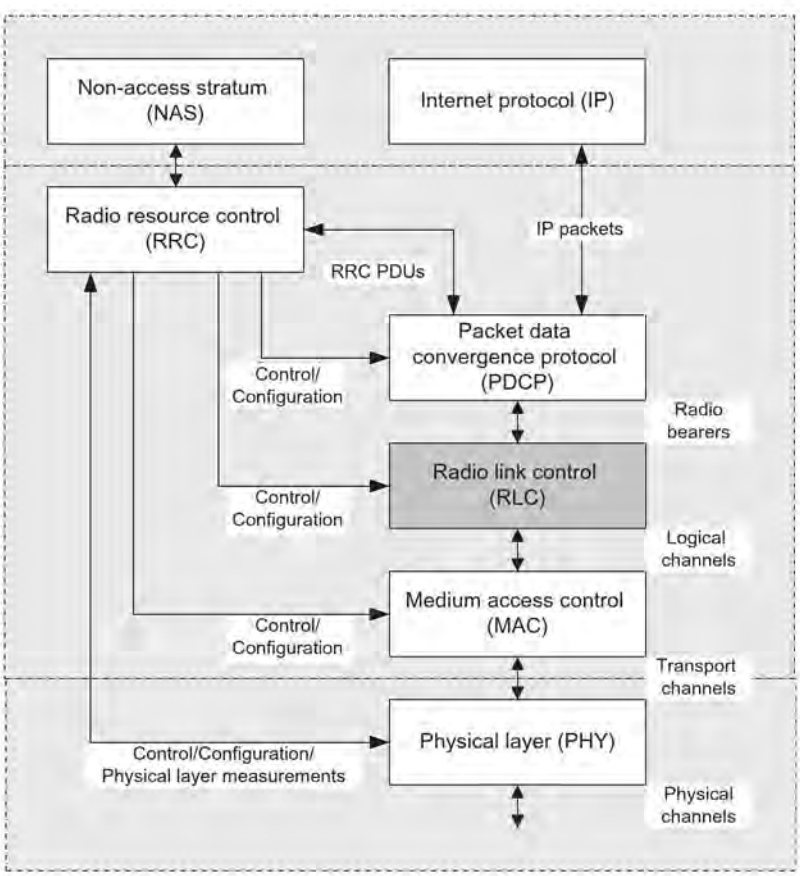


FIGURE 7.1

The RLC sublayer in the LTE protocol stack.

The main services and functions of the RLC sublayer include transfer of upper layer PDUs, error correction through automatic repeat request (ARQ) (AM data transfer), concatenation, segmentation and reassembly of RLC SDUs (UM and AM data transfer), resegmentation of RLC data PDUs (AM data transfer), reordering of RLC data PDUs (UM and AM data transfer), duplicate detection (UM and AM data transfer), protocol error detection (AM data transfer), RLC SDU discard (UM and AM data transfer), and RLC reestablishment [8].

The segmentation and concatenation functions are meant to generate RLC PDUs of appropriate size from the incoming RLC SDUs. One option would be to define a fixed PDU size; however, if the chosen size is too large, it will not be possible to efficiently support small-packet applications and low data rates. Moreover, excessive padding would be required in some scenarios. On the other hand, a single small PDU size would result in an excessive overhead due to the the protocol header included in each PDU. To avoid these disadvantages, which are especially important given the large variety of data applications and services supported by LTE, the RLC PDU size is designed to dynamically vary. The RLC header includes information fields such as sequence number (SN), which is used by the reordering and retransmission mechanisms to sort the PDUs. The reassembly function at the receiver side performs the reverse operation to reassemble the SDUs from the received PDUs [11].

Retransmission of missing PDUs is one of the main functions of the RLC sublayer. Although most of the errors can be detected and corrected by the HARQ protocol at the physical layer and MAC sublayer [1,5,7], there are benefits in having a complementary retransmission mechanism such as ARQ. By inspecting the SNs of the received PDUs, missing PDUs can be detected and a retransmission can be requested from the transmitting entity. Different services have different requirements, i.e., for some services (e.g., FTP), error-free delivery of data is essential, whereas for other applications (e.g., streaming services and VoIP), a small amount of missing packets will not cause a significant degradation.

In-sequence delivery implies that data blocks are delivered by the receiver in the same order as they were transmitted. This is an essential task of the RLC sublayer. The HARQ processes operate independently and the transport blocks may be delivered out of sequence to the RLC sublayer. In-sequence delivery implies that SDU n should be delivered prior to SDU $n + 1$. This is an important feature, as many applications require the data to be received in the same order as it was transmitted. As an example, TCP transport protocol can limitedly handle IP packets arriving out of sequence with some performance degradation; however, for some streaming applications in-sequence delivery is critical. The main idea behind in-sequence delivery is to store the received PDUs in a buffer until all PDUs with lower SN have been delivered. In practice, the RLC retransmission module, when operating in the acknowledged mode, operates on the same buffer as the in-sequence delivery mechanism [11].

The RLC PDUs are exchanged with the MAC sublayer, and the data transfer between the two protocol layers in the uplink and downlink is through logical channels. The RLC PDUs are of variable size and can be formatted based on the transport block size available to the MAC sublayer considering the physical channel conditions. The MAC sublayer notifies the RLC sublayer when a transmission opportunity becomes available, including the size of the RLC PDUs that can be transmitted in the current transmission opportunity.

In this chapter, we will take a top-down systematic approach to describe various functions and services of the LTE/LTE-Advanced RLC sublayer, starting with the RLC architecture and then analyzing each function and service separately to provide an in-depth understanding of this E-UTRA protocol layer.

7.1 RLC architecture

The functions of the RLC sublayer are controlled by the RRC sublayer and are performed by RLC entities. For each RLC entity configured at the eNB, there is a corresponding RLC entity configured at the UE. As shown in Figure 7.2, an RLC entity receives RLC SDUs from the upper layer and sends RLC PDUs to its peer RLC entity via lower layers. An RLC PDU can either be an RLC data PDU or an RLC control PDU. In the transmitter side, the RLC entity receives RLC SDUs from the upper layer through a single SAP between the RLC and the upper layer, and after forming the RLC data PDUs from the received RLC SDUs, the RLC entity delivers the RLC data PDUs to the lower layer through a single logical channel. Similarly, in the receiver side the RLC entity receives RLC data PDUs from a lower layer, through a single logical channel, and after constructing the RLC SDUs from the received RLC data PDUs, the RLC entity delivers the RLC SDUs to the upper layer through a single SAP between the RLC and the upper layer. Depending on the link direction, an RLC entity delivers or receives RLC control PDUs to or from lower layers through the same logical channel that it delivers or receives the RLC data PDUs.

The RLC sublayer receives data from upper layer radio bearers (signaling and data) called SDUs. The transmission entities in the RLC sublayer convert them to RLC PDUs after performing functions such as segmentation, concatenation, and adding RLC headers. In the other direction, the receiving entities receive RLC PDUs from the MAC sublayer. After performing reordering, the RLC PDUs are reassembled into RLC SDUs and are delivered to the upper layer. For each mode, there is a transmitting entity and a receiving entity. The transmitting and receiving entities operate independently for TM and UM. For AM, the transmitting and receiving entities operate together since retransmissions have to be performed. Therefore, feedback received by the receiving entity is given to the transmitting entity to make retransmission decisions.

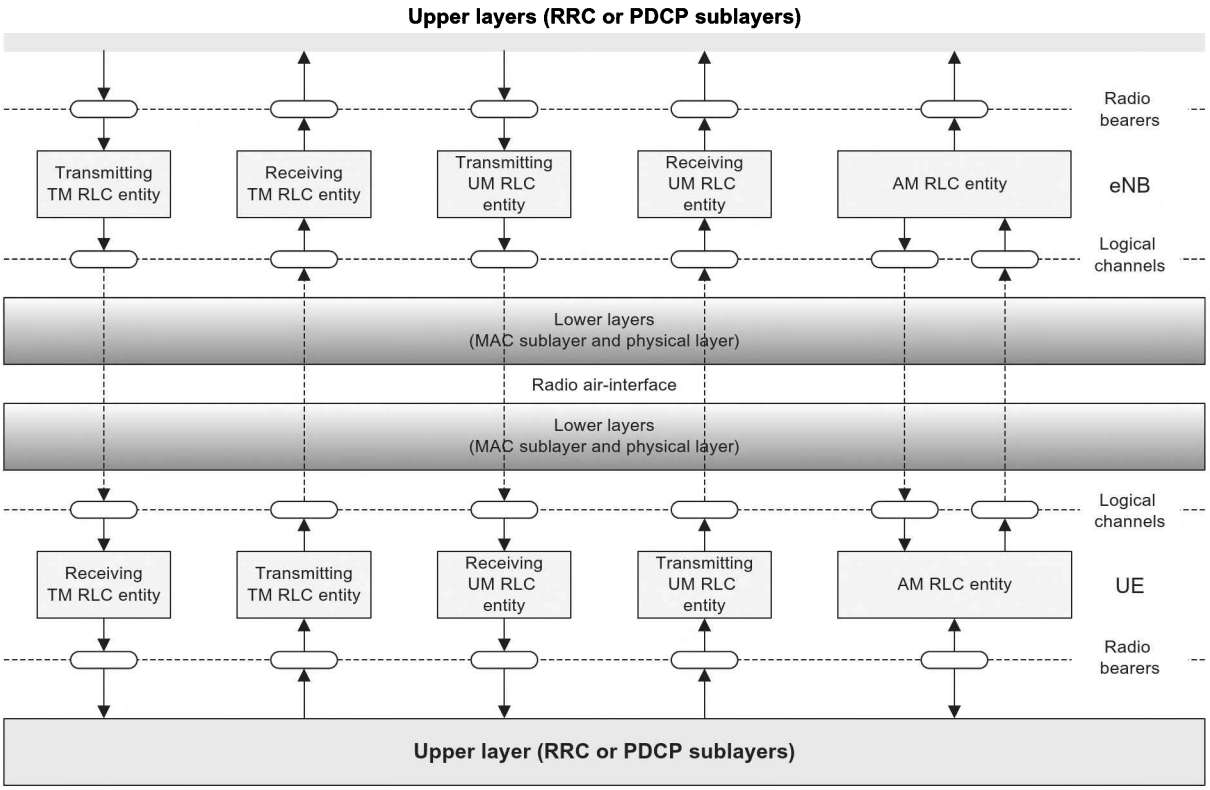


FIGURE 7.2

RLC sublayer architecture [2].

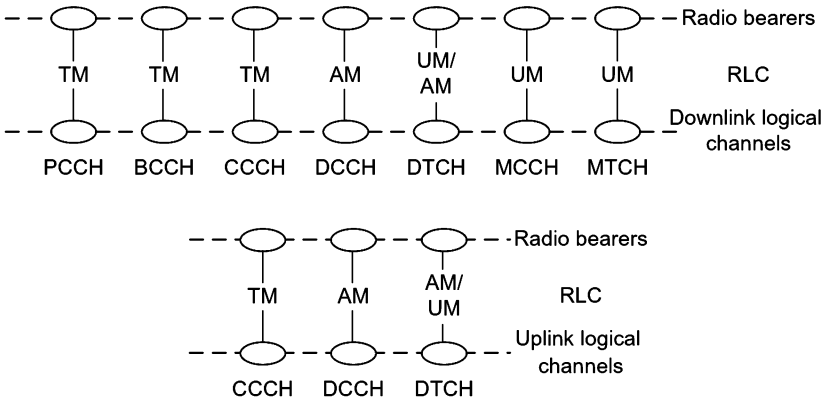


FIGURE 7.3

Downlink and uplink RLC channel mapping.

As mentioned earlier, an RLC entity can be configured to perform data transfer in TM, UM, or AM modes. As a result, an RLC entity is categorized as a TM RLC entity, a UM RLC entity, or an AM RLC entity depending on the mode of data transfer that the RLC entity is configured to operate. As shown in Figure 7.2, a TM/AM/UM RLC entity is configured either as a transmitting TM/AM/UM RLC entity or a receiving TM/AM/UM RLC entity (note that the operation modes must match on both sides). The transmitting RLC entity receives RLC SDUs from the upper layer and sends RLC PDUs to its peer receiving RLC entity via lower layers. The receiving RLC entity delivers RLC SDUs to the upper layer and receives RLC PDUs from its peer transmitting RLC entity via lower layers [2].

Figure 7.3 shows the RLC mode association with different logical channels in the downlink. The TM usage is limited to common control signaling channels of BCCH, PCCH, and CCCH. In LTE, the UM is applied only to traffic channels. The AM is applicable to traffic as well as dedicated control channels. Figure 7.3 also shows RLC mode association with different logical channels in the uplink where only AM and UM are applicable. The UM is utilized for common control signaling in the form of CCCH and can be additionally applied to traffic channels. The AM is applicable to traffic as well as dedicated control channels.

7.2 RLC transfer modes

As we discussed earlier in this chapter, each radio bearer in the LTE is processed in one of the following three transfer modes:

1. **Transparent Mode:** This mode can be regarded as null RLC since it is simply a pass through. None of the major RLC functions are applicable to this

mode. The RLC sublayer does not add any header or other overhead. The use of TM is limited to the common signaling channels such as BCCH, PCCH, and CCCH.

2. **Unacknowledged Mode:** This mode provides all RLC functions except retransmissions and resegmentation; therefore providing unreliable service. It would typically be employed for real-time applications such as VoIP, which are error tolerant and delay sensitive, but still require other functions such as in-sequence delivery and duplicate detection.
3. **Acknowledged Mode:** This mode provides a reliable transmission by offering error correction and retransmissions. It provides the entire RLC functions and is suited for applications that require reliability such as TCP/IP packet data services (e.g., browsing or e-mail).

The octet aligned RLC SDUs of variable sizes are supported for all RLC entity types (i.e., TM, UM, and AM RLC entities). The RLC PDUs are formed upon notification of a transmission opportunity and are delivered to the MAC sublayer. In the following sections, we discuss RLC functions associated with each of the above transfer modes in more detail. Table 7.1 summarizes the RLC functions that are supported by each RLC transfer mode [2].

7.2.1 Transparent mode

The TM entity consists of simply a transmission buffer to hold the RLC SDUs until a transmission opportunity becomes available at the lower layers. No other processing is done by the transmitting TM RLC entity. The receiving TM RLC entity passes the received PDUs to the higher layers. The TM entity does not segment or concatenate RLC SDUs. Therefore, each RLC SDU is an RLC PDU. The framing and identification of the SDU boundary must be performed by the upper layers. As shown in Figure 7.4, a TM RLC entity can be configured to deliver/receive RLC PDUs through BCCH, downlink/uplink CCCH, and PCCH logical channels. When a transmitting TM RLC entity forms transparent mode data

Table 7.1 Support of RLC Functions in Each Mode [2]

RLC Functions	TM	UM	AM
Transfer of upper layer PDUs	Yes	Yes	Yes
Error correction through ARQ	No	No	Yes
Concatenation, segmentation, and reassembly of RLC SDUs	No	Yes	Yes
Resegmentation of RLC data PDUs	No	No	Yes
Reordering of RLC data PDUs	No	Yes	Yes
Duplicate detection	No	Yes	Yes
RLC SDU discard	No	Yes	Yes
RLC reestablishment	Yes	Yes	Yes
Protocol error detection	No	No	Yes

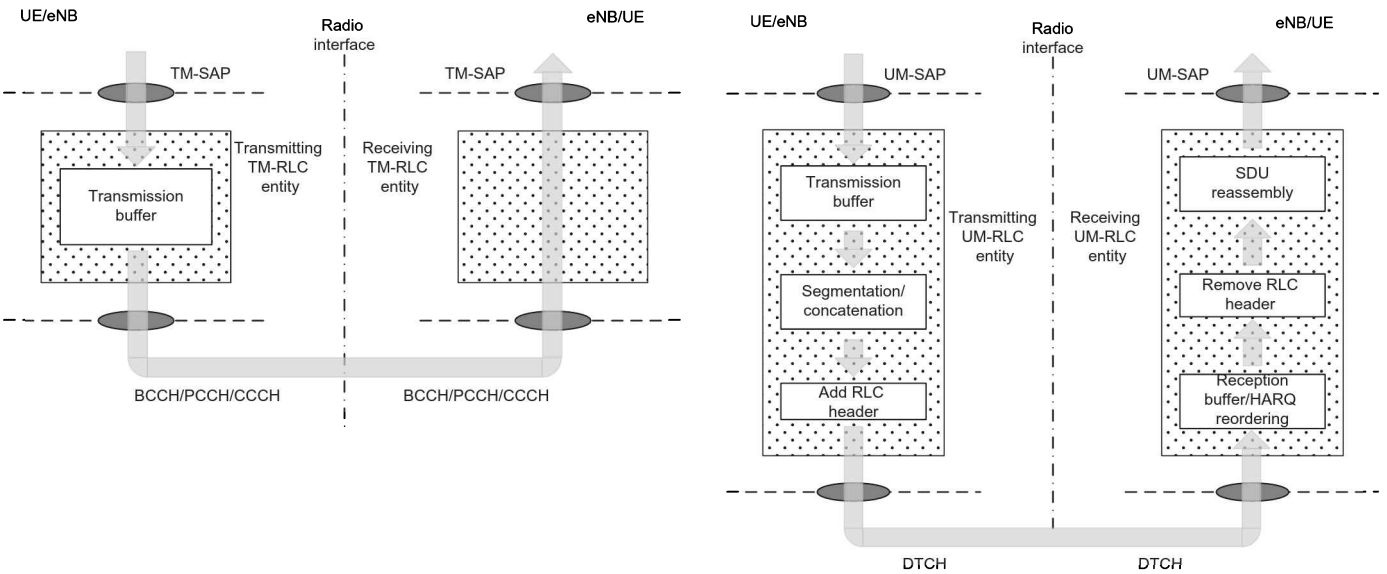


FIGURE 7.4

Illustration of LTE RLC TM and UM entities [2].

(TM data) PDUs from RLC SDUs, it does not segment or concatenate the RLC SDUs, and does not include any RLC headers in the TM data PDUs [2,8].

7.2.2 Unacknowledged mode

A UM RLC entity can be configured to deliver/receive RLC PDUs through downlink/uplink DTCH logical channels. When a transmitting UM RLC entity constructs unacknowledged mode data (UM data) PDUs from RLC SDUs, it segments and/or concatenates the RLC SDUs (so that the UM data PDUs can match the total size of RLC PDU specified by the lower layer protocol at a particular transmission opportunity), and includes relevant RLC headers in the UM data PDU. When a receiving UM RLC entity receives UM data PDUs, it detects whether duplicate UM data PDUs have been received and discards duplicate PDUs, it reorders the UM data PDUs, if they are received out of sequence, it detects the loss of UM data PDUs at lower layers and avoids excessive reordering delays, it reassembles RLC SDUs from the reordered UM data PDUs and delivers the RLC SDUs to the upper layer protocol in ascending order of the RLC SN, and it discards received UM data PDUs that cannot be reassembled into an RLC SDU due to loss of a UM data PDU which belonged to a particular RLC SDU at lower layers. At the time of RLC reestablishment, the receiving UM RLC entity reassembles RLC SDUs from the UM data PDUs that are received out of sequence and delivers them to the upper layer protocol, discards any remaining UM data PDUs that could not be reassembled into RLC SDUs, initializes relevant state variables, and stops relevant timers [2].

As shown in Figure 7.4, the UM transmitting entity places the received RLC SDUs in the transmit buffer. When a transmission opportunity becomes available, it may perform segmentation or concatenation of RLC SDUs depending on SDU size as well as the size of transmission opportunity. After segmentation/concatenation, an RLC header is added. The RLC header includes information fields such as SN and length. The resulting RLC PDU is passed onto the MAC sublayer for transmission. The UM receiving entity holds received UM RLC PDUs in the receive buffer. The PDUs may be out of order due to HARQ retransmissions. Therefore, they are reordered based on their SNs. After removing RLC headers, the data fields of RLC PDUs are assembled back into SDUs, undoing any segmentation and concatenation, and are delivered to the upper layers.

Figure 7.5 shows an example segmentation/concatenation operation on RLC SDUs. In this example, the nth RLC SDU is segmented with the first part transmitted as part of the PDU with SN = 1, and the second part transmitted in the PDU with SN = 2. Two additional complete SDUs ($n + 1$, $n + 2$) are concatenated in this PDU. The last SDU $n + 3$ is segmented with the first segment included as the last data field in the PDU with SN = 2 and the other segment is included in the next PDU with SN = 3. For the PDU with SN = 2, the header is set accordingly with the framing info (FI) set to "11" to indicate that the RLC PDU includes segments of SDUs at both ends. Multiple length indicators (LI) are

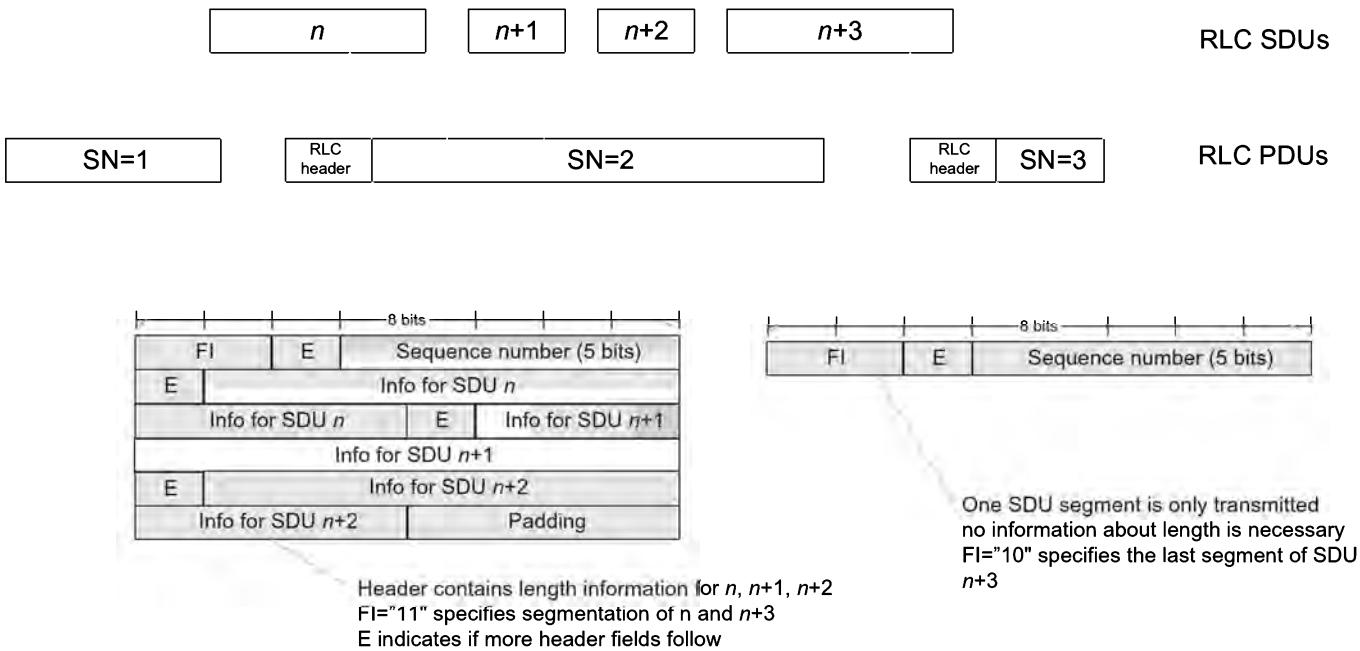


FIGURE 7.5

Illustration of UM segmentation and concatenation functions.

included. Each LI identifies the length of one complete or partial SDU included in the PDU except the last one. For the PDU with SN = 3, no extension fields are necessary as it contains only data from one SDU. In this case, the FI field is set to "10" to indicate that the last byte of the PDU data corresponds to the last byte of the SDU.

Figure 7.6 illustrates an example UM reordering process, which is controlled by parameters *UM_Window_Size* and *T_reordering*. The reordering window specifies the number of PDUs that can be received without advancing the window. If the window advances, any PDUs outside the window, regardless of their status, are delivered to the upper layers. The reordering timer determines how long the receiving entity waits for a missing PDU to be transmitted. It is started as soon as a sequence gap is detected. Upon expiration of the timer, the receiving entity will deliver any PDUs with contiguous packets before or after the sequence gap to the upper layers. In this example, the transmitting entity sends PDUs 0 through 3. PDUs 0 and 3 are lost, while PDUs 1 and 2 are successfully received. The receiving entity starts the *T_reordering* timer once the gap due to PDU 0 is detected. However, before the expiration of *T_reordering*, PDU 0 is delivered (out of sequence) following a HARQ retransmission, thus contiguous PDUs 0, 1, and 2 are delivered to the upper layers. The reordering window is then advanced. Next, the transmitter sends three more PDUs (4, 5, and 6) which are successfully received. The *T_reordering* timer is now started due to the sequence gap caused by PDU 3. Finally, *T_reordering* expires, causing the receiver to stop waiting for PDU 3, and deliver contiguous PDUs beyond PDU 3 (PDUs 4, 5, and 6) to the upper layers.

7.2.3 Acknowledged mode

As shown in Figure 7.7, an AM RLC entity can be configured to deliver or receive RLC PDUs through downlink or uplink DCCH/DTCH logical channels. When the transmitting side of an AM RLC entity constructs acknowledged mode data (AM data) PDUs from RLC SDUs, it segments and/or concatenates the RLC SDUs so that the AM data PDUs fit within the total size of the RLC PDU at a particular transmission opportunity. The transmitting side of an AM RLC entity supports retransmission of RLC data PDUs (i.e., ARQ). If the RLC data PDU that needs to be retransmitted does not fit within the total size of the RLC PDU at a particular transmission opportunity, the AM RLC entity can resegment the RLC data PDU into AM data PDU segments where the number of resegmentations is not limited [2].

In general, when the transmitting side of an AM RLC entity forms AM data PDUs from RLC SDUs received from an upper layer or AM data PDU segments from RLC data PDUs to be retransmitted, it includes relevant RLC headers in the RLC data PDU. The receiving side of an AM RLC entity receives RLC data PDUs and detects and discards the duplicate RLC data PDUs. It reorders the RLC data PDUs, if they are received out of sequence, and detects the missing RLC data

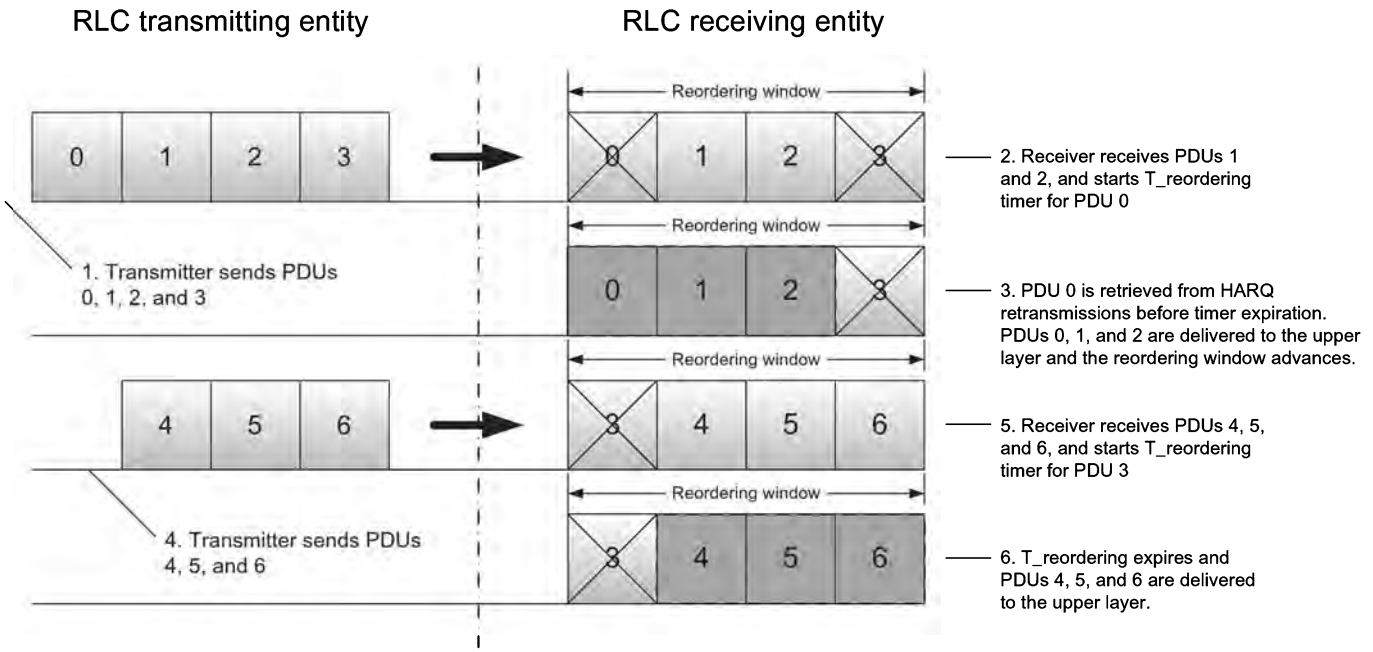


FIGURE 7.6

Illustration of the UM reordering process.

PDUs at the lower layers and sends retransmission requests to its peer AM RLC entity. It further reassembles RLC SDUs from the reordered RLC data PDUs and delivers the RLC SDUs to the upper layer in sequence. At the time of RLC reestablishment, the receiving side of an AM RLC entity reassembles RLC SDUs from the RLC data PDUs that are received out of sequence and delivers them to the upper layer, discarding the remaining RLC data PDUs that could not be reassembled into RLC SDUs, initializing relevant state variables, and stopping relevant timers.

The transmitting AM RLC entity places the received RLC SDUs in the transmission buffer. When a transmission opportunity becomes available, the SDUs in the transmission buffer are segmented or concatenated depending on the permissible size of the packet. An RLC header is added to each PDU prior to passing them to the MAC sublayer for transmission. The RLC PDUs are also placed in the retransmission buffer in case retransmission is necessary. When the receiver

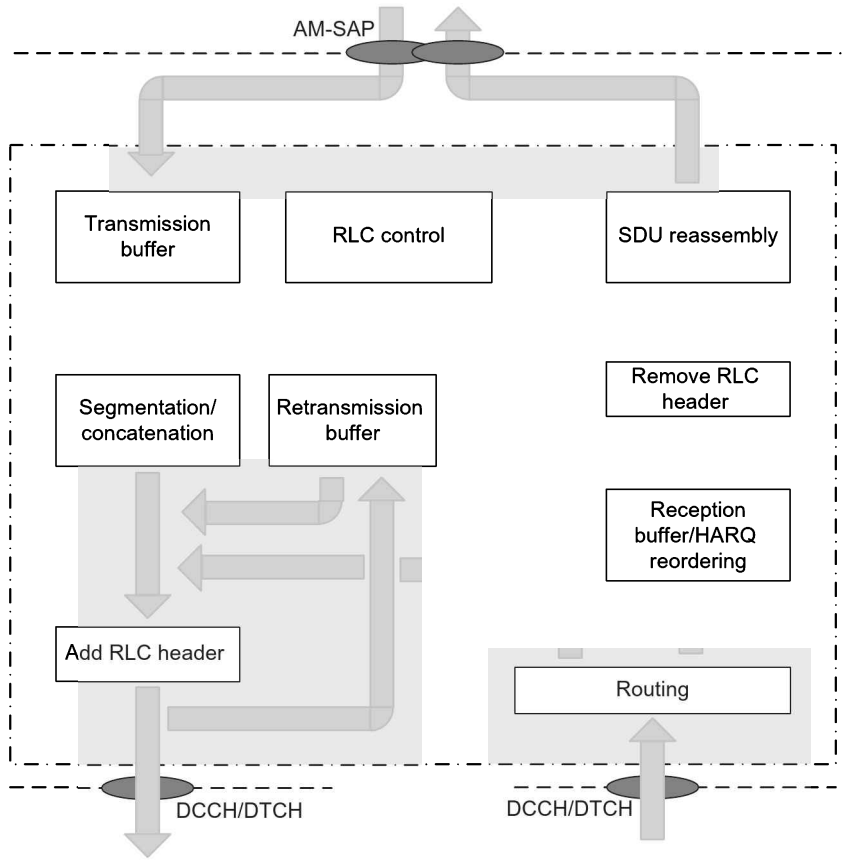


FIGURE 7.7

Illustration of LTE RLC AM entity [2].

sends an ACK or NACK, the transmitting entity attempts a retransmission. If an ACK is received for the entire RLC PDU, then that PDU is flushed from the retransmission buffer. If an NACK is received for part of a PDU or the entire PDU, the transmitting entity schedules a retransmission. If the size of transmission opportunity does not allow the entire RLC PDU to be re-sent, then resegmentation is possible whereby a single PDU is divided into multiple segments. Each segment can then be transmitted as a separate PDU with the RLC header indicating how the segmentation was carried out. Otherwise, the entire RLC PDU is scheduled for retransmission [2].

The receiving RLC entity accumulates the received RLC PDUs in the receive buffer, and then it performs the reordering before passing the complete SDU to the higher layers. Status PDUs are sent by the receiving entity to acknowledge the received PDUs and to signify missing PDUs or parts of missing PDU segments. Figure 7.8 shows an example illustration of the formation of an RLC AM status PDU. The status report shows that the receiving entity has received PDUs up to (but not including) SN = 8. Moreover, *NACK_SN* fields are included to indicate that PDUs with SN = 0, 3, and 6 are entirely missing. A *NACK_SN* is included for the PDU with SN = 5, but this is accompanied by *SOstart* and *SOend* to indicate that only a portion of the PDU between bytes 50 and 99 is missing.

The status PDU is sent from the receiving AM RLC entity to indicate successfully received as well as missing PDUs or portions of missing PDUs assuming resegmentation was used. The status PDU is sent when the poll bit is set in a PDU sent from the transmitting RLC entity while timer *T_status_prohibit* has expired, or *T_reordering* expires (indicating a missing PDU) and *T_status_prohibit* has already expired. *T_status_prohibit* is a timer that ensures duplicate status PDUs are not sent unnecessarily, and is set based on the RLC round trip time. It is restarted whenever a status PDU is sent. The transmitting AM entity sets the poll bit to request the receiving AM entity to send status PDUs. The frequency of the poll bit inclusion is controlled by the following parameters or events: (1) if the number of

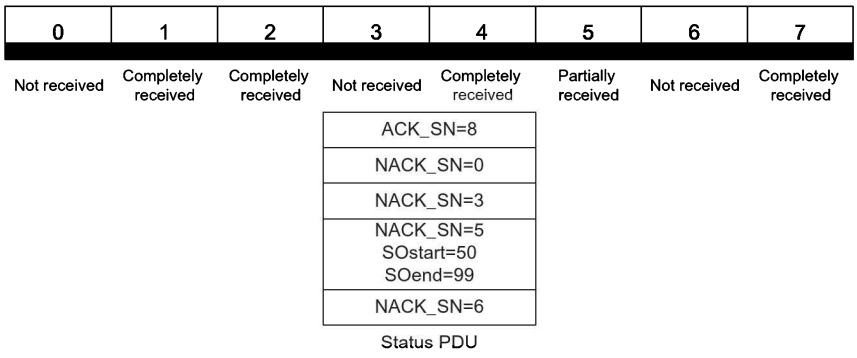


FIGURE 7.8

Illustration of an RLC AM status PDU.

PDUs sent since the last status report is received exceeds *Poll_PDU*; (2) if the total number of bytes of the PDUs sent since receiving the last status report exceeds *Poll_Byte*; (3) if the *T_poll_retransmit* timer expires and there are no PDUs to be sent in the transmission or retransmission buffer, the transmitting AM RLC entity resends the last PDU which was sent with poll bit set; (4) if no PDUs can be sent, the poll bit is set in the last PDU. This trigger is important in LTE since variable size RLC PDUs are supported [2].

In RLC AM, both transmit and receive entities maintain a window for flow control and reordering purposes. Figure 7.9 illustrates the mechanism for advancing the transmit/receive windows, as well as the behavior of the receiving RLC entity when encountering sequence gaps (missing packets) in the data stream. The sequence gaps (missing packets) would be filled using HARQ or RLC retransmissions that are triggered by status reporting. For simplicity, the length of transmit

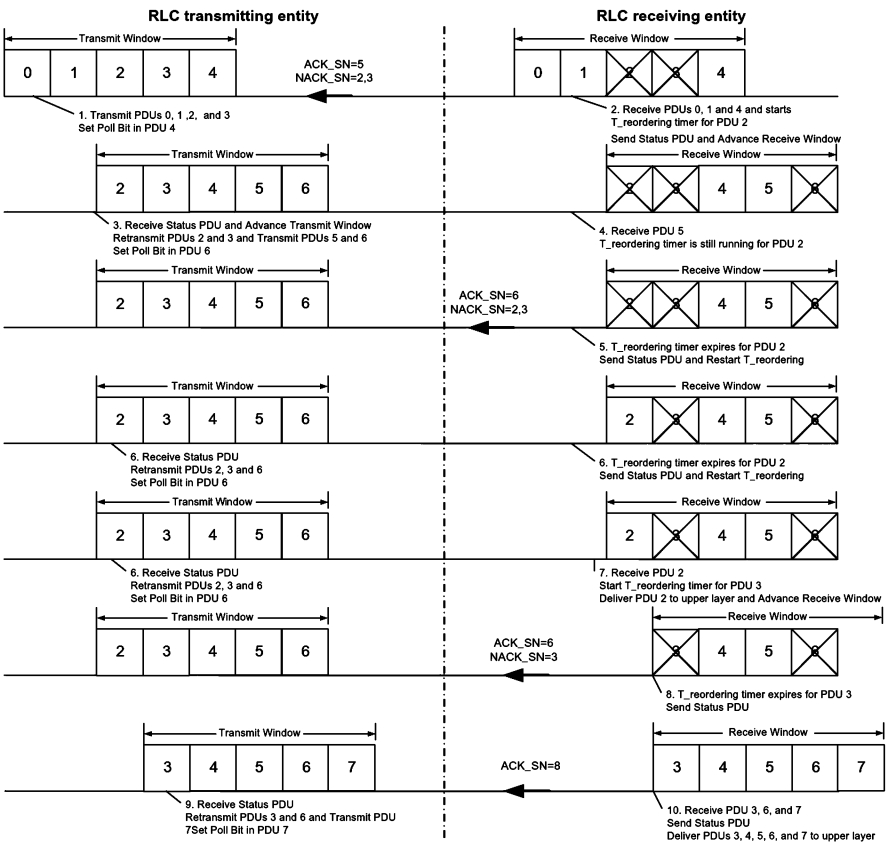


FIGURE 7.9

Illustration of an RLC AM reordering process.

and receive windows in this example is equal to 5 packets, whereas in reality the actual size of the RLC AM window is 512.

Based on the example shown in Figure 7.9, the process of RLC AM reordering can be described using the following steps:

1. Five PDUs are put in the transmit window for transmission.
2. The receiver receives PDUs 0, 1, and 4, but does not receive PDUs 2 and 3. PDUs 0 and 1 are delivered to the higher layers. A status PDU is sent indicating the next not-received PDU is SN = 5 and the missing PDUs are 2 and 3. The *T_reordering* timer is started, since a gap as result of the missing PDU is created within the receive window.
3. The transmitter retransmits the missing PDUs 2 and 3, and transmits new PDUs 5 and 6 as the transmit window can be advanced by two PDUs.
4. The receiver does not receive PDUs 2, 3, and 6, but receives PDU 5. The *T_reordering* timer associated with PDU 2 is still running.
5. The *T_reordering* timer associated with PDU 2 expires. This triggers the transmission of another status PDU, informing the transmitter that the next not-received PDU is 6, and that gaps exist due to PDUs 2 and 3.
6. After receiving the status PDU, the transmitter retransmits PDUs 2, 3, and 6.
7. The receiver successfully receives PDU 2, delivers it to the upper layers and advances the receive window. The *T_reordering* timer is then started for PDU 3.
8. The *T_reordering* timer expires as PDU 3 was never received. This causes the receiver to transmit a status PDU informing the transmitter that the next not-received PDU is 6, and that a gap exists due to PDU 3.
9. The transmitter receives the status PDU and retransmits PDUs 3 and 6, and transmits PDU 7 as the transmit widow can now be advanced.
10. The receiver successfully receives PDUs 3, 6, and 7. A status PDU is transmitted notifying the transmitter that all PDUs up to and including PDU 7 have been received.

The RLC sublayer in LTE has the ability to resegment previously transmitted PDUs. This enables the RLC sublayer to create segments whose size is suited to the transmission opportunity reported by the MAC sublayer. The example illustrated in Figure 7.10 shows the process of resegmentation with the following steps:

1. The transmitter sends an RLC PDU with SN = 101 with a length of 100 bytes.
2. The receiver sends a status PDU containing *NACK_SN* = 101.
3. As the indicated transmission opportunity now will only permit a transmission of 25 bytes, the RLC decides to resegment the PDU. The transmitted SN is still 101, but now the segment field is included (SO = 0) indicating that this is a resegmented PDU, and this is the first part of the original PDU. The LSF field is set to 0 to indicate that this is not the last segment.
4. The transmitter sends the second segment (25 bytes) of the RLC PDU with SO = 25. The LSF is still set to 0.

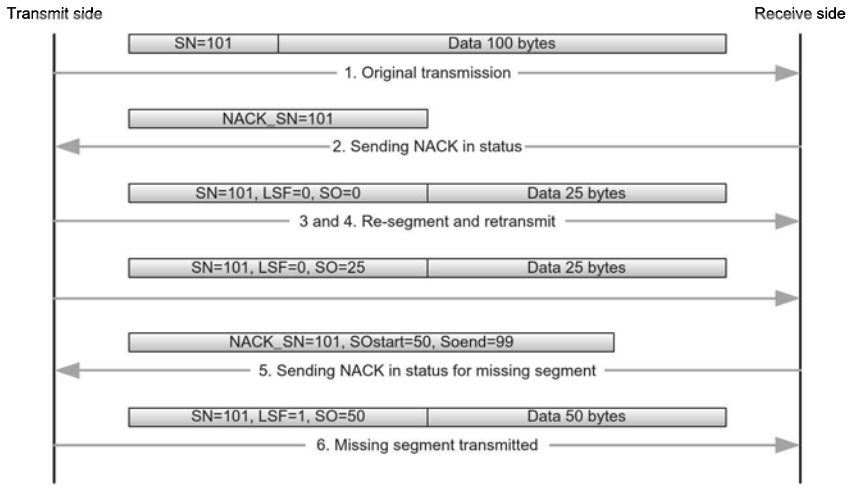


FIGURE 7.10

Illustration of an RLC AM resegmentation process.

5. The receiver sends a status PDU with $NACK\_SN = 101$. It also includes $SOstart = 50$ and $SOend = 99$ to indicate the portion of the RLC PDU that has not been received yet.
6. The transmitter sends the third segment (50 bytes) of the RLC PDU with SO = 50. The LSF is set to 1 to indicate that this is the last segment of the RLC PDU. The receiver may send a status PDU, acknowledging the RLC PDU, if it has received all segments correctly.

7.3 RLC ARQ procedure

The ARQ is an error detection and correction mechanism that is incorporated in the AM RLC entity to improve the reliability and robustness of packet transmission. In the following sections, we briefly describe the theory of ARQ and later explain the ARQ procedures and relevant state variables and timers [12,13].

7.3.1 ARQ principles

ARQ is an error-control mechanism for data transmission which uses acknowledgments (ACK) (or negative acknowledgments (NACK)) and timeouts to achieve reliable data transmission over an unreliable communication link. In an ARQ scheme, the receiver uses an error detection code, typically a cyclic redundancy check (CRC), to detect if the received packet is in error [6,12,13]. If no error is detected in the received data, the transmitter is notified by sending a positive ACK. If an error is detected, the receiver discards the packet and sends a

NACK to the transmitter and requests a retransmission. An ACK or a NACK is a short message sent by the receiver to the transmitter to indicate whether it has correctly or incorrectly received a data packet, respectively. Timeout is a predetermined time interval after the sender sends the packet; if the sender does not receive an ACK before the timeout, it usually retransmits the packet until it receives an ACK or exceeds a predefined number of retransmissions. There are three types of ARQ protocol as follows [12,13]:

1. **Stop-and-Wait ARQ** is the basic form of ARQ protocol where the sender sends one packet at a time and then waits for an ACK or NACK signal from the receiver before sending the same or a new packet. The receiver sends an ACK signal following receipt of a good packet. If the ACK does not reach the sender before the timeout, the sender resends the same packet.
2. **Go-Back-N ARQ** is a form of ARQ protocol in which the sender continuously sends a number of packets (determined by the duration of transmission window) without receiving an ACK signal from the receiver. The receiver process keeps track of the sequence number of the next packet it expects to receive and sends the sequence number with every ACK it sends. The receiver will ignore any packet that does not have the exact sequence number it expects, whether that packet is a duplicate of a packet it has already acknowledged or a packet with a sequence number higher than the one expected. Once the sender has sequentially sent all the packets in its transmission window, it will check whether all of the packets are acknowledged and will resume sequential transmission of the packets starting with the next sequence number to the one that was last acknowledged.
3. **Selective Repeat ARQ** is a form of ARQ protocol for the transmission and acknowledgment of packets or fragments of a packet where the sending process continues to send a number of packets specified by a window size even after a packet is lost. Unlike Go-Back-N ARQ, the receiving process will continue to accept and acknowledge packets sent after an initial error. The receiver process keeps track of the sequence number of the earliest packet it has not received, and sends that number with every acknowledgment it sends. If a packet from the sender does not reach the receiver, the sender continues to send subsequent packets until it has exhausted its window. The receiver continues to fill its receiving window with the subsequent packets, replying each time with an acknowledgment containing the sequence number of the earliest missing packet. Once the sender has sent all the packets in its window, it resends the packet number given by the acknowledgments, and then continues where it left off.

Figure 7.11 provides an example illustration of the ARQ variants. It is shown how efficiently different ARQ schemes utilize the communication channel and use the ARQ buffers in the transmitter and receiver. The major advantage of ARQ over forward error correction (FEC) schemes is that error detection requires much simpler decoding mechanisms and much less redundancy than error correcting codes. Furthermore, ARQ is adaptive in the sense that information is

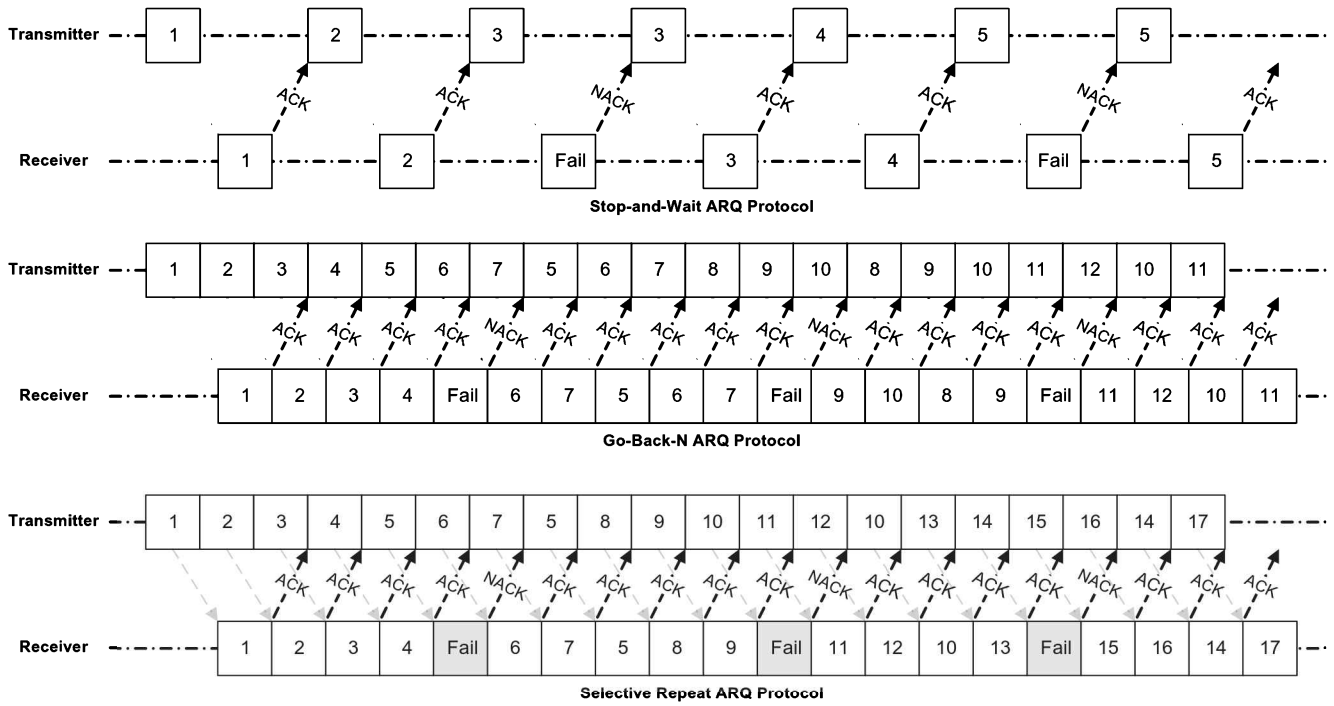


FIGURE 7.11

Illustration of different ARQ schemes [12,13].

retransmitted only when error occurs. On the other hand, FEC may be desirable instead of, or in addition to, error detection, for any of the following reasons: (1) a feedback channel is not available or ARQ delay is not tolerable; (2) the retransmission scheme is not conveniently implemented; and (3) the expected number of errors without correction would require excessive retransmissions [12]. The ARQ can be used for each connection between the UE and the eNB. Since the use of ARQ may increase the latency due to more reliable transmission, the ARQ mechanism is usually disabled for delay-sensitive applications such as VoIP or interactive gaming.

7.3.2 ARQ operation

The ARQ processes in the transmitter and receiver entities maintain separate state machines to track the transmission, retransmission, and discarding of the packets exchanged between the two entities. The state machines on each side employ several state variables and counters that are described in this section. The transmitting side of each AM RLC entity maintains the following state variables [2]:

1. Acknowledgment state variable *VT(A)*: This state variable holds the value of the sequence number of the next AM data PDU for which a positive acknowledgment will be received in-sequence, and it serves as the lower edge of the transmitting window. It is initially set to zero and is updated whenever the AM RLC entity receives a positive acknowledgment for an AM data PDU with $SN = VT(A)$.
2. Maximum send state variable *VT(MS)*: This state variable equals to *VT(A)* + *AM_Window_Size*, and it serves as the higher edge of the transmitting window.
3. Send state variable *VT(S)*: This state variable holds the value of the *SN* to be assigned for the next newly generated AM data PDU. It is initially set to zero and is updated whenever the AM RLC entity delivers an AM data PDU with $SN = VT(S)$.
4. Poll send state variable *POLL_SN*: This state variable holds the value of *VT (S)*-1 upon the most recent transmission of an RLC data PDU with the poll bit set to one. It is initially set to zero.

The transmitting side of each AM RLC entity maintains the following counters [2]:

1. *PDU_WITHOUT_POLL* counter: This counter is initially set to zero. It counts the number of AM data PDUs sent since the most recent poll bit was transmitted.
2. *BYTE_WITHOUT_POLL* counter: This counter is initially set to zero. It counts the number of data bytes sent since the most recent poll bit was transmitted.
3. *RETX_COUNT* counter: This counter counts the number of retransmissions of an AM data PDU. There is one *RETX_COUNT* counter per PDU that needs to be retransmitted.

The receiving side of each AM RLC entity maintains the following state variables [2]:

1. Receive state variable *VR(R)*: This state variable holds the value of the *SN* following the last in-sequence completely received AM data PDU, and it serves as the lower edge of the receiving window. It is initially set to zero and is updated whenever the AM RLC entity receives an AM data PDU with $SN = VR(R)$.
2. Maximum acceptable receive state variable *VR(MR)*: This state variable equals to $VR(R) + AM\_Window\_Size$, it holds the value of the *SN* of the first AM data PDU that is beyond the receiving window, and serves as the higher edge of the receiving window.
3. *t-Reordering* state variable *VR(X)*: This state variable holds the value of the *SN* following the *SN* of the RLC data PDU which triggered *t-Reordering*.
4. Maximum status transmit state variable *VR(MS)*: This state variable holds the highest possible value of the *SN* which can be indicated by *ACK_SN* when a status PDU needs to be constructed. It is initially set to zero.
5. Highest received state variable *VR(H)*: This state variable holds the value of the *SN* following the *SN* of the RLC data PDU with the highest *SN* among received RLC data PDUs. It is initially set to zero.

Figure 7.10 illustrates the application of some of the ARQ state variables and timers in RLC AM. In the process of transmission of packets between peer RLC entities, the transmitting side of an AM RLC entity may receive an NACK, which is a notification of reception failure by its peer AM RLC entity, for an AM data PDU or part of an AM data PDU through a status PDU received from its peer AM RLC entity. In this case, if the sequence number of the corresponding AM data PDU falls within the range $VT(A) \leq SN < VT(S)$, the AM data PDU or part of the AM data PDU for which a NACK was received is retransmitted.

When retransmitting a portion of an AM data PDU, the transmitting side of an AM RLC entity divides the AM data PDU into a number of segments, where each segment forms a new AM data PDU segment which can fit within the size of RLC PDU permissible by the lower layer at a particular transmission opportunity and delivers the new AM data PDU segment to the lower layer.

An AM RLC entity can poll its peer AM RLC entity in order to trigger status reporting at the peer AM RLC entity. Upon assembly of an AM data PDU or AM data PDU segment, the transmitting side of an AM RLC entity includes a poll in the RLC data PDU, when the transmission buffer and the retransmission buffer become empty after transmission of the RLC data PDU, or when no new RLC data PDU can be transmitted after the transmission of the last RLC data PDU.

An AM RLC entity sends status PDUs to its peer AM RLC entity in order to provide positive and/or negative acknowledgments for RLC PDUs. The RRC signaling determines whether the status-prohibit function can be used for an AM RLC entity. The triggers to initiate status reporting include polling from a peer AM RLC entity (this ensures that the RLC status report is transmitted after

HARQ reordering), as well as detection of reception failure of an RLC data PDU (the receiving side of an AM RLC entity triggers a status report when *t-Reordering* expires). A status PDU can be triggered when polling from a peer entity is received or reception failure of RLC data PDU is detected (*t-Reordering* timer expiration). As long as the *t-StatusProhibit* timer is running, a status PDU cannot be sent to the lower layer. When a status PDU is sent to the lower layer, *t-StatusProhibit* is restarted. Note that the *t-StatusProhibit* timer is configured by the upper layer [2].

When constructing a status PDU, the AM RLC entity must include an *NACK_SN* which is set to the *SN* of the AM data PDU for the AM data PDUs with sequence numbers satisfying $VR(R) \leq SN < VR(MS)$ that have not been completely received, in increasing *SN* order of PDUs and increasing byte segment order within PDUs, starting with $SN = VR(R)$ up to the point where the resulting status PDU would still fit the total size of the RLC PDU(s) permissible by the lower layer protocol, or for an AM data PDU for which no byte segments have been received. For an uninterrupted sequence of byte segments of a partly received AM data PDU that have not been received, a set of *NACK_SN*, *SOstart*, and *SOend* are included in the status PDU. The *ACK_SN* is set to the *SN* of the next not-received RLC data PDU, which is not indicated as missing in the resulting status PDU [2].

7.4 RLC PDU packet formats

The RLC sublayer provides TM, UM, and AM data transfer services including indication of successful delivery of PDUs to the RRC or PDCP sublayers. The MAC sublayer provides data transfer services to the RLC sublayer, as well as notification of a transmission opportunity (along with the total size of the RLC PDUs). The RLC PDUs can be categorized into RLC data and control PDUs. RLC data PDUs are used by TM, UM, and AM RLC entities to transfer upper layer PDUs. RLC control PDUs are used by AM RLC entity to perform ARQ procedures. The status PDU is used by the receiving side of an AM RLC entity to inform the peer AM RLC entity about correctly received, as well as lost, RLC data PDUs. Figure 7.12 provides the high-level aspects of packet processing in LTE layer 2 protocols, by illustrating IP packet processing stages in various sublayers.

The RLC SDUs are bit strings that are octet aligned. An RLC SDU is included into an RLC PDU starting from its first bit onward. In the following sections, we discuss the formats of the RLC PDUs when operating in different data transfer modes.

7.4.1 Transparent mode PDU format

The TM data PDU only consists of a data field and does not include any RLC headers. The structure of RLC TM data PDUs is shown in Figure 7.13.

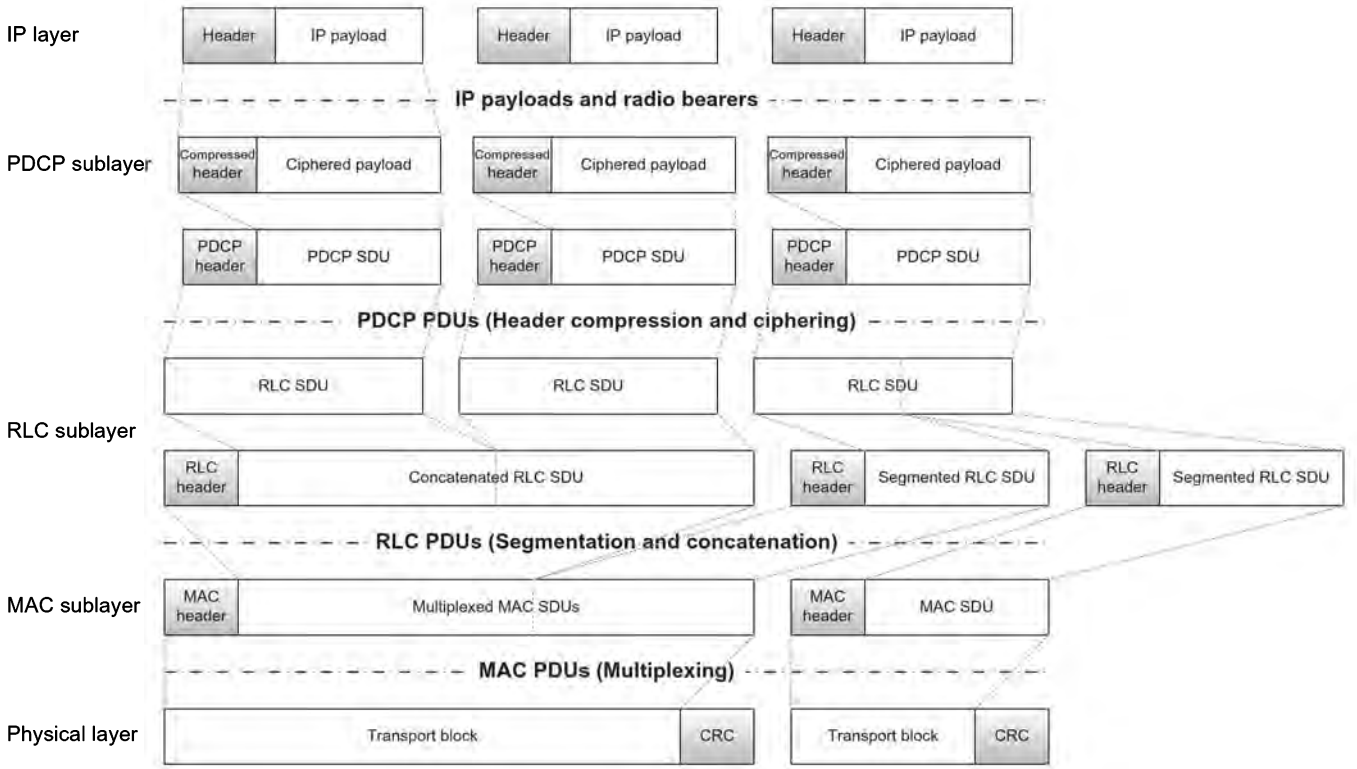


FIGURE 7.12

IP packet processing in different LTE layer 2 sublayers [8,9].

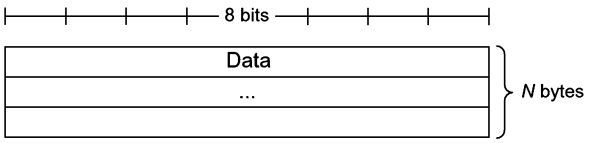


FIGURE 7.13

TM data PDU format [2,9].

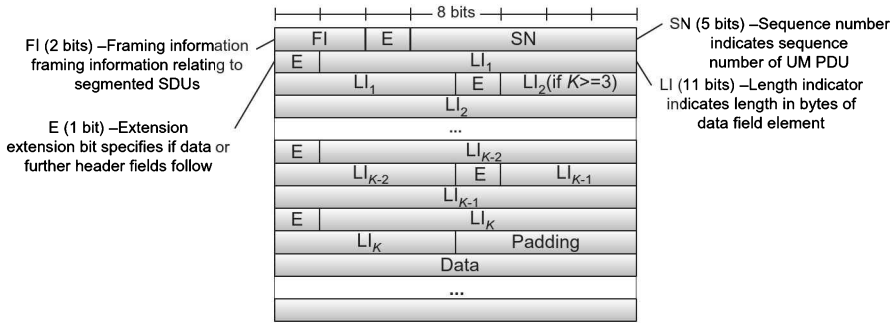


FIGURE 7.14

Unacknowledged mode data PDU format [2,9].

7.4.2 Unacknowledged mode PDU formats

The UM data PDU formats for a single and partial SDU are shown in Figure 7.14. There are different formats based on the length of SN, which may be 5 or 10 bits. The framing info (FI) field indicates whether a PDU contains a segment of RLC SDU either at the beginning or at the end. When multiple SDUs are concatenated into a single PDU, header extension fields are used to specify the size of each SDU with an 11-bit length indicator (LI) specified for each SDU except the last one. A single-bit extension (E) indicator specifies whether more header fields follow each LI. The LI for the last data field is not necessary as the size of the PDU will be signaled by the MAC layer. The UM data PDU header consists of a fixed part, i.e., information fields that are present for every UM data PDU, and an extension part, i.e., information fields that are present for a UM data PDU when necessary. The fixed part of the UM data PDU header is octet aligned and consists of FI, E, and SN fields. The extension part of the UM data PDU header is octet aligned and consists of E(s) and LI(s) fields. A UM data PDU header includes an extension part only when more than one data field elements are present. In that case, E and LI fields are present for every data field element except the last one. Furthermore, when a UM data PDU header consists of an odd number of LI(s), four padding bits will follow after the last LI field [2].

Samsung Ex. 1005, Page 365 of 1147

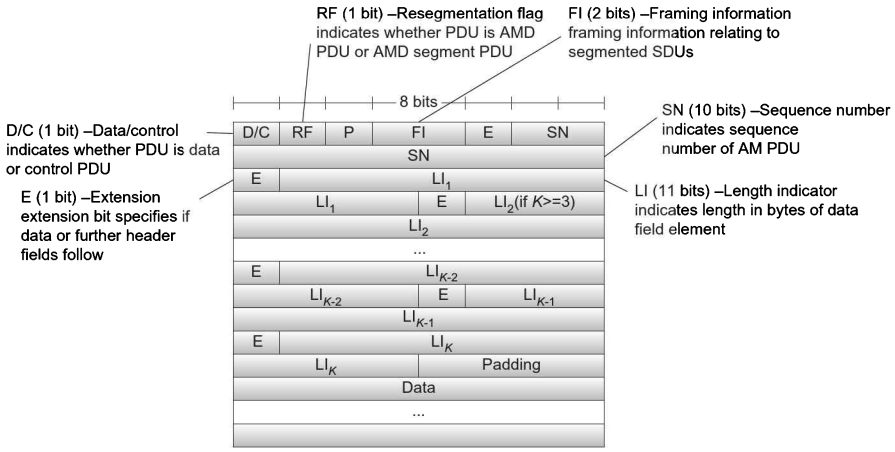


FIGURE 7.15

Acknowledged mode data PDU format [2,9].

7.4.3 Acknowledged mode PDU formats

Figure 7.15 shows the PDU format for transmitting AM data PDUs for the first time, or retransmitting AM data PDUs without resegmentation. The RLC AM transfer mode always operates with 10-bit SN. There is a one-bit D/C field that indicates whether the PDU carries data bits from upper layers or is a control PDU. The resegmentation flag (RF) is a one-bit field that indicates whether this PDU contains retransmitted PDUs that have been segmented. For first time transmissions, this flag is set to zero. The polling bit (P) is set whenever the transmitting AM entity requests that the receiving AM entity sends a status PDU. The FI field indicates whether parts of the RLC SDU are included in the PDU either at the beginning or at the end, and the interpretation is identical to the UM case. Similarly, the extension bit (E) indicates whether LI fields are included in the PDU. If multiple RLC SDUs are concatenated in the RLC PDU, the E bit and the LI are included for each SDU. The E bit indicates whether this is the last LI field before a data field [2].

The AM data PDU resegmentation format is used when an RLC PDU is resegmented for retransmission. The RLC header is identical to the normal RLC data PDU with the exception of the following differences (Figure 7.16):

- Re-segmentation flag (RF): Set to one to indicate resegmentation has been done.
- Last segment flag (LSF): The LSF field indicates whether this AM PDU segment contains the last segment of the retransmitted RLC PDU.
- Segment offset (SO): This 15-bit field indicates the position of the segment within the original RLC PDU. Specifically, the SO field indicates the position within the data field of the original AM data PDU to which the first byte of the data field of the AM data PDU segment belongs.

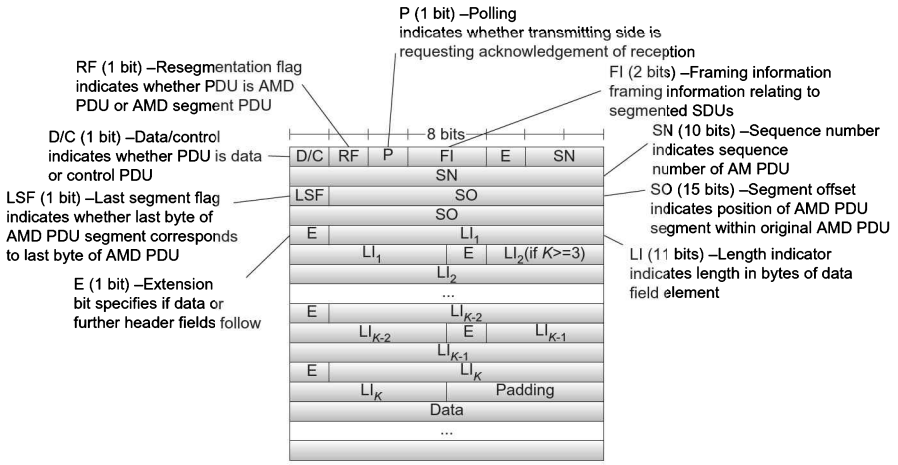


FIGURE 7.16

Acknowledged mode data PDU segment format [2,9].

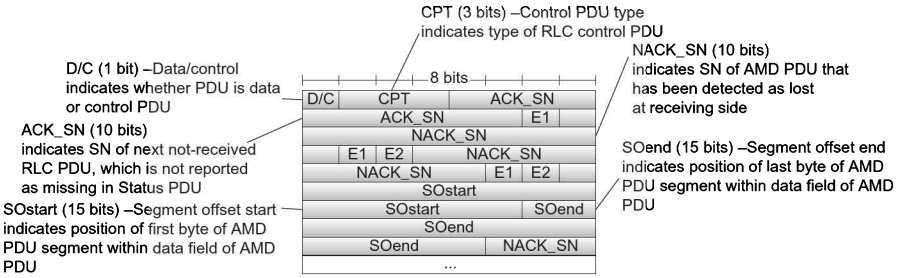


FIGURE 7.17

Status PDU format [2,9].

7.4.4 Status PDU format

The status PDU is sent from the receiving AM entity to the transmitting AM entity to provide information about received and missing RLC PDUs. The structure of status PDU is shown in Figure 7.17. The status PDU can also indicate missing segments of a PDU. The status PDU contains the following fields [2]:

- **D/C:** Data or control flag is set to one to indicate a control PDU.
- **ACK_SN:** The SN of the PDU up to which the receiver has received all the PDUs excluding those indicated by NACK_SN field(s).
- **NACK_SN:** Indicates the SN of a PDU which is not completely or only partially received. NACK_SN indicates gaps in the sequence of PDUs acknowledged by the ACK_SN.

Table 7.2 RLC Constants and Timers [2]

Timers and Constants	Description and Usage	Value
T_poll_retransmit	Used by transmitting side of AM RLC entity to retransmit a poll	0 to 200 ms
T_reordering	Used by receiving side of AM RLC entity and UM RLC entity to detect loss of RLC PDUs at the lower layer	0 to 500 ms
T_status_prohibit	Used by receiving side of AM RLC entity to prohibit transmission of a status PDU	0 to 500 ms
AM_Window_Size	Used by both transmitting side and receiving side of each AM RLC entity to calculate higher edge of window from lower edge	512
UM_Window_Size	Used by receiving UM RLC entity to define SNs of those UM data PDUs that can be received without causing an advancement of the receiving window	16 for 5-bit SN 512 for 10-bit SN

- **E1:** Extension field which indicates whether a set of NACK_SN, E1, and E2 fields follow.
- **E2:** Extension field which indicates whether SOstart and SOend fields are included following NACK_SN.
- **SOstart and SOend:** Indicate the start and stop positions within a segmented PDU which has not been successfully received. The SN is specified by the preceding NACK_SN.

7.5 RLC variables, constants, and timers

You may have noticed that the overall data flow of the RLC sublayer seems relatively easy to understand. But understanding the RLC sublayer mechanism in such detail that would help with troubleshooting and performance optimization in a real implementation would not be a simple task. One of the challenges is to intimately comprehend the definition and usage of the state variables, timers, and counters.

The state variables related to the RLC AM data transfer mode can take values from 0 to 1023. All arithmetic operations on state variables related to AM data transfer are adjusted by the AM modulus defined as *final_value* = [value from arithmetic operation]mod1024. The state variables related to the RLC UM data transfer mode can take values from 0 to $2^{(\text{sn-FieldLength})} - 1$. All arithmetic operations on state variables related to UM data transfer are adjusted by the UM modulus defined as *final_value* = [value from arithmetic operation]mod$2^{(\text{sn-FieldLength})}$. The AM and UM

Table 7.3 Description of RLC Transfer Mode Parameters [2]

Parameter	Category	Description
VT(A)	State variable	ACK state variable This state variable holds the value of the *SN* of the next AM data PDU for which a positive ACK is going to be received in-sequence, and it serves as the lower edge of the transmitting window It is initially set to zero and is updated whenever the AM RLC entity receives a positive ACK for an AM data PDU with *SN* = *VT(A)*
VT(MS)	State variable	Maximum send state variable *VT(MS)* = *VT(A)* + *AM_Window_Size* It serves as the higher edge of the transmitting window
VT(S)	State variable	Send state variable This state variable holds the value of the *SN* to be assigned for the next newly generated AM data PDU. It is initially set to zero and is updated whenever the AM RLC entity delivers an AM data PDU with *SN* = *VT(S)*
VR(R)	State variable	Receive state variable This state variable holds the value of the *SN* following the last in-sequence completely received AM data PDU. It serves as the lower edge of the receiving window. It is initially set to zero and is updated whenever the AM RLC entity receives an AM data PDU with *SN* = *VR(R)*
VR(MR)	State variable	Maximum acceptable receive state variable *VR(MR)* = *VR(R)* + *AM_Window_Size* It holds the value of the *SN* of the first AM data PDU that is beyond the receiving window and serves as the higher edge of the receiving window
VR(X)	State variable	*t-Reordering* state variable This state variable holds the value of the *SN* following the *SN* of the RLC data PDU which triggered *t-Reordering*
VR(MS)	State variable	Maximum status transmit state variable This state variable holds the highest possible value of the *SN* which can be indicated by *ACK_SN* when a status PDU needs to be constructed. It is initially set to zero

(*Continued*)

Table 7.3 (Continued)

Parameter	Category	Description
VR(H)	State variable	Highest received state variable This state variable holds the value of the *SN* following the *SN* of the RLC data PDU with the highest *SN* among received RLC data PDUs. It is initially set to zero
VT(US)	State variable	This state variable holds the value of the *SN* to be assigned for the next newly generated UM data PDU. It is initially set to zero and is updated whenever the UM RLC entity delivers a UM data PDU with *SN* = *VT(US)*
VR(UR)	State variable	UM receive state variable This state variable holds the value of the SN of the earliest UM data PDU that is still considered for reordering. It is initially set to zero
VR(UX)	State variable	UM *t-Reordering* state variable This state variable holds the value of the *SN* following the *SN* of the UM data PDU which triggered t-Reordering
VR(UH)	State variable	UM highest received state variable This state variable holds the value of the *SN* following the *SN* of the UM data PDU with the highest *SN* among received UM data PDUs, and it serves as the higher edge of the reordering window. It is initially set to zero
POLL_SN	State variable	Poll send state variable This state variable holds the value of *VT(S)*-1 upon the most recent transmission of an RLC data PDU with the poll bit set to one. It is initially set to zero
PDU_WITHOUT_POLL	Counter	This counter is initially set to zero. It counts the number of AM data PDUs sent since the most recent poll bit was transmitted
BYTE_WITHOUT_POLL	Counter	This counter is initially set to zero. It counts the number of data bytes sent since the most recent poll bit was transmitted
RETX_COUNT	Counter	This counter counts the number of retransmissions of an AM data PDU. There is one *RETX_COUNT* counter per PDU that needs to be retransmitted
AM_Window_Size	Constant	*AM_Window_Size* = 512
UM_Window_Size	Constant	*UM_Window_Size* = 16 when *SN* = 5 bits *UM_Window_Size* = 512 when *SN* = 10 bits *UM_Window_Size* = 0 when UM RLC is for MTCH or MCCH

(*Continued*)

Table 7.3 (Continued)

Parameter	Category	Description
t-PollRetransmit	Timer	Used by the transmitting side of an AM RLC entity in order to retransmit a poll
t-Reordering	Timer	This timer is used by the receiving side of an AM RLC entity and receiving UM RLC entity in order to detect loss of RLC PDUs at the lower layer. Only one t-Reordering per RLC entity is running at a given time
t-StatusProhibit	Timer	This timer is used by the receiving side of an AM RLC entity in order to prohibit transmission of a status PDU
maxRetxThreshold	Configurable parameters	This parameter is used by the transmitting side of each AM RLC entity to limit the number of retransmissions of an AM data PDU
pollPDU	Configurable parameters	This parameter is used by the transmitting side of each AM RLC entity to trigger a poll for every *poll PDU* PDUs
pollByte	Configurable parameters	This parameter is used by the transmitting side of each AM RLC entity to trigger a poll for every poll bytes
sn-FieldLength	Configurable parameters	This parameter gives the UM *SN* field size in bits

data PDUs are numbered with integer-valued SNs in the range of 0 to 1023 for AM data PDU and 0 to $2^{(\text{sn-FieldLength})} - 1$ for UM data PDU. A summary of RLC constants and timers as well as state parameters are provided in Tables 7.2 and 7.3.

References

3GPP Specifications

[1] 3GPP TS 36.321, Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol Specification (Release 11), December 2012.
[2] 3GPP TS 36.322, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) Protocol Specification (Release 11), December 2012.
[3] 3GPP TS 36.323, Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) Specification (Release 11), December 2012.
[4] 3GPP TS 36.331, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC) Protocol Specification (Release 11), December 2012.
[5] 3GPP TS 36.211, Evolved Universal Terrestrial Radio Access (E-UTRA), Physical Channels and Modulation (Release 11), December 2012.

[6] 3GPP TS 36.212, Evolved Universal Terrestrial Radio Access (E-UTRA) Multiplexing and Channel Coding (Release 11), December 2012.
[7] 3GPP TS 36.213, Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures (Release 11), December 2012.
[8] 3GPP TS 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Overall description, Stage 2 (Release 11), December 2012.

Books

[9] C. Johnson, Long Term Evolution in Bullets, second ed., Create Space Independent Publishing Platform, July 2012.
[10] S. Sesia, I. Toufik, M. Baker, LTE, the UMTS Long Term Evolution: From Theory to Practice, second ed., John Wiley & Sons, August 2011.
[11] E. Dahlman, S. Parkvall, J. Skold, 4G: LTE/LTE-Advanced for Mobile Broadband, Academic Press, May 2011.
[12] S. Lin, D. Costello, Error Control Coding, second ed., Prentice Hall, June 2004.

Articles

[13] R. Comroe, D. Costello, ARQ schemes for data transmission in mobile radio systems, IEEE J. Sel. Areas Commun. 2 (4) (July 1984).

Samsung Ex. 1005, Page 372 of 1147

This page intentionally left blank

CHAPTER

Medium Access Control Functions

8

CHAPTER CONTENTS

8.1 MAC architecture and services 345
8.2 Logical and transport channel mapping 349
8.3 Downlink/uplink data transfer 353
8.3.1 HARQ principles 354
8.3.2 Downlink HARQ timing and signaling protocols 357
8.3.3 Uplink HARQ timing and signaling protocols 358
8.4 DRX operation 366
8.5 MAC control elements 371
8.5.1 Uplink timing advance control element 372
8.5.2 BSR control element 374
8.5.3 Power headroom control element 375
8.5.4 Extended power headroom control element 376
8.5.5 C-RNTI control element 377
8.5.6 Activation/deactivation control element 378
8.5.7 MCH scheduling information control element 378
8.5.8 UE contention resolution identity control element 379
8.6 MAC PDU packet formats 380
8.7 MAC scheduling services 381
8.7.1 Dynamic and semi-persistent scheduling 384
8.7.2 Proportional fair scheduling algorithm 386
8.8 RNTI mapping 387
8.9 Random access procedure 389
8.9.1 Contention-based RACH procedure 392
8.9.2 Contention-free RACH procedure 394
References 396

LTE-Advanced.

© 2014 Elsevier Inc. All rights reserved.

Samsung Ex. 1005, Page 374 of 1147

The layer 2 in LTE are classified into medium access control (MAC), radio link control (RLC), and packet data convergence protocol (PDCP) functions [2,3]. The MAC sublayer is responsible for data transfer and logical channel multiplexing, hybrid automatic repeat request (HARQ) retransmissions and uplink/downlink scheduling, random access procedure, and maintenance of uplink timing. It is also responsible for multiplexing/demultiplexing data across multiple component carriers when carrier aggregation is used. From a physical layer (PHY) perspective, the MAC sublayer provides and receives services in the form of transport channels. A transport channel is defined by how and with what characteristics the information is transmitted over the radio interface. The data in a transport channel is organized into transport blocks. In each transmission time interval (TTI), one or two transport blocks of dynamic size are transmitted over the radio interface to a terminal depending on whether spatial multiplexing is supported. Each transport block is associated with a transport format, specifying how the transport block is physically processed and transmitted over the radio interface. The transport format includes information about the transport block size, the modulation and coding scheme (MCS), and the multi-antenna mapping [5,6]. By varying the transport format, the MAC sublayer can realize different data rates and reliability. Figure 8.1 illustrates the MAC sublayer in the LTE protocol stack. The interaction of the MAC sublayer with other layers in the protocol stack, as well as the service access points with the physical layer and RLC sublayers, are shown in the figure.

The functions of the MAC sublayer in LTE can be more systematically classified as follows:

- Channel mapping: MAC sublayer maps logical channels carrying RLC protocol data units (PDUs) to transport channels. For transmission, multiple service data units (SDUs) from logical channels are mapped to the transport block carried over transport channels. In the receiving end, transport blocks from transport channels are demultiplexed and assigned to the corresponding logical channels.
- Scheduling: MAC sublayer performs all scheduling-related functions in the uplink and downlink, and thus it is responsible for transport format selection associated with all transport channels. This includes HARQ functionality. Since the scheduling services are implemented at the eNB, the MAC sublayer is responsible for reporting scheduling-related information, such as UE buffer status and power headroom. It also handles prioritization from both an inter-UE and intra-UE logical channel perspective.
- Random access procedures: MAC sublayer is responsible for random access procedures in the uplink that can be performed following a contention or non-contention-based process.
- Uplink timing maintenance: UE needs to maintain timing synchronization with the cell at all times. The MAC sublayer performs required procedures for periodic synchronization.

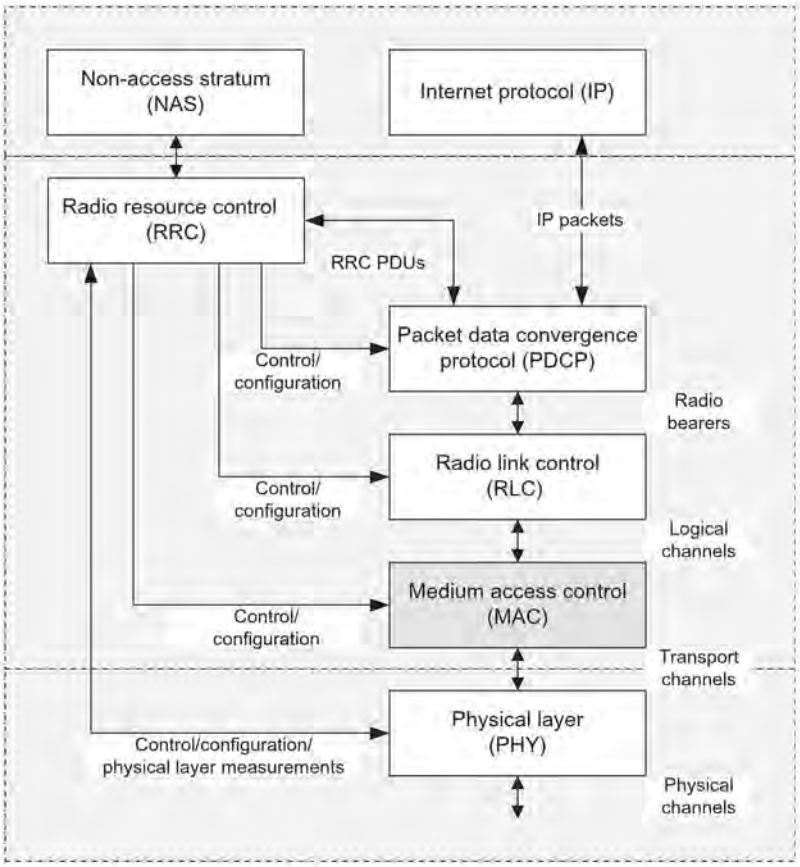


FIGURE 8.1

The MAC sublayer in the LTE protocol stack.

Figure 8.2 illustrates the IP packet processing in different parts of LTE layer 2. As shown in the figure, the MAC sublayer receives the RLC SDUs mapped to various logical channels in the downlink and generates the MAC PDUs that further become the transport blocks in the physical layer. In the following sections, we take a top-down systematic approach to study and understand various MAC sublayer services in LTE/LTE-Advanced and how those functions interact and interoperate with other protocol layers.

8.1 MAC architecture and services

The architecture of the LTE MAC sublayer is shown in Figure 8.3. It comprises a HARQ entity, a multiplexing/demultiplexing entity, a logical channel

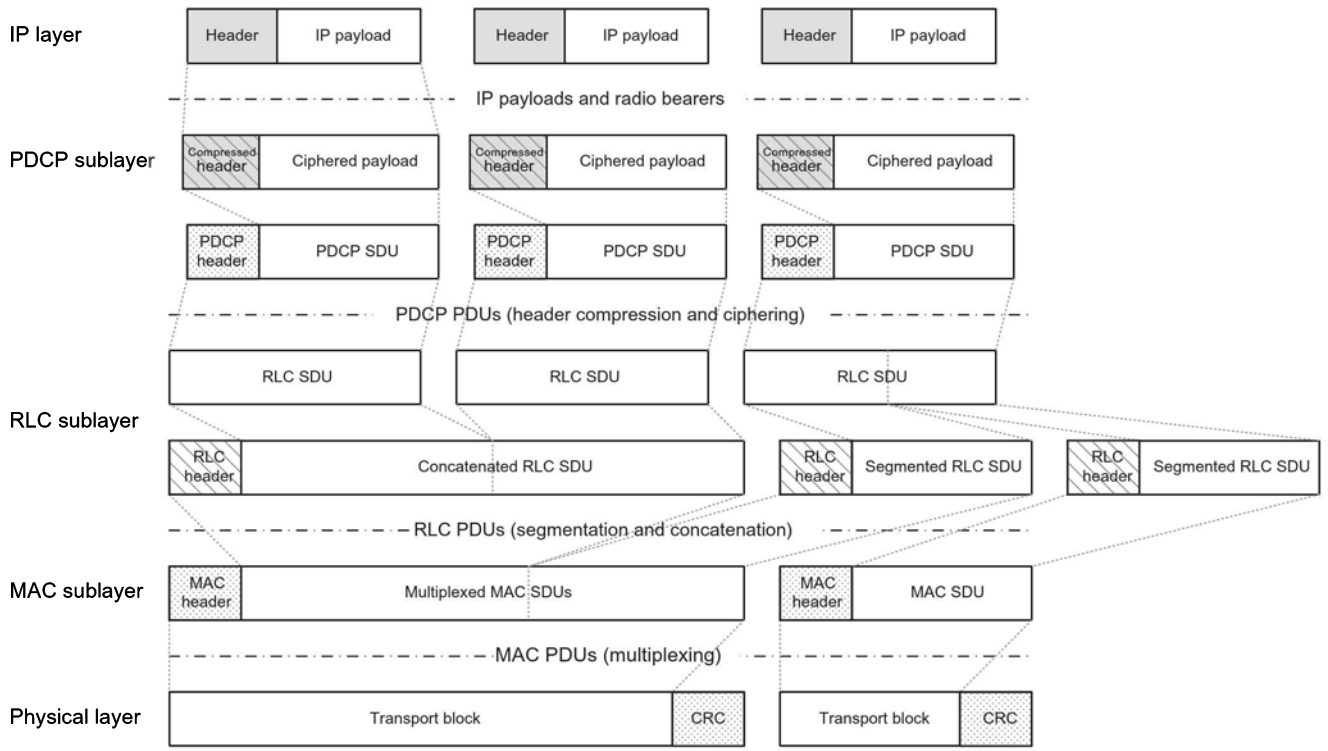


FIGURE 8.2

IP packet processing in different sublayers of LTE Layer 2 [8].

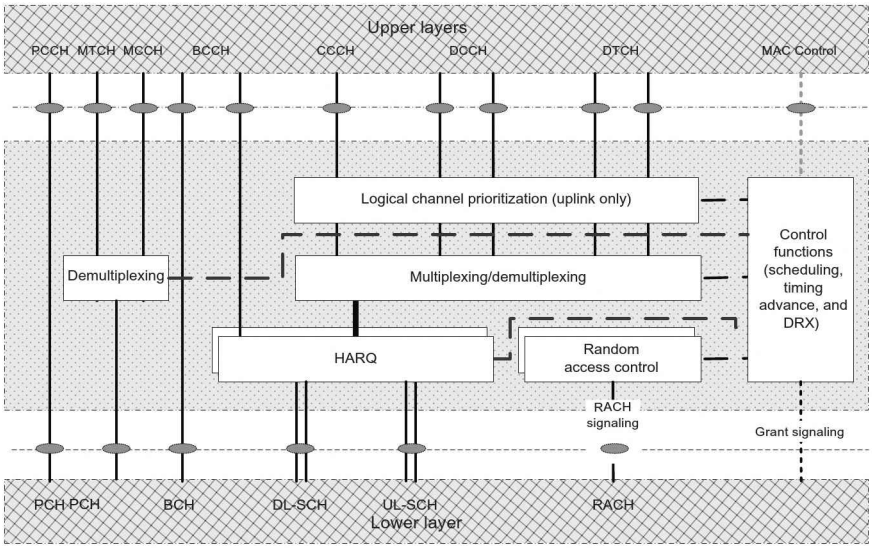


FIGURE 8.3

LTE UE MAC sublayer functions and services [1].

prioritization entity, a random access control entity, and a controller which performs various control functions, such as scheduling, timing advance maintenance, and discontinuous reception (DRX). The HARQ entity is responsible for downlink/uplink HARQ operations. The transmit HARQ operation includes transmission and retransmission of transport blocks, and receiving and processing of acknowledgment/negative-acknowledgment (ACK/NACK) signaling. The receive HARQ operation includes reception of transport blocks, combining of the received data and generation of ACK/NACK signaling. In order to enable continuous transmission while previous transport blocks are being decoded, up to eight HARQ processes in parallel are used to support a multi-process stop-and-wait HARQ protocol, which implies that in each HARQ process, upon transmission of a transport block, the transmitter stops further transmissions and waits for feedback from the receiver. When a NACK is received, or when a certain time period elapses without receiving any feedback, the transmitter retransmits the transport block. Each HARQ process is responsible for a separate stop-and-wait operation and manages a separate buffer.

In general, HARQ protocols can be categorized as synchronous or asynchronous, and the retransmissions in each case can be adaptive or non-adaptive. In a synchronous HARQ scheme, the retransmission(s) for each process occur at predefined time intervals relative to the initial transmission time. In this method, there is no need to signal information, such as the HARQ process number, as this can be inferred from the transmission timing. In contrast, in an asynchronous HARQ scheme, the retransmissions can occur at any time relative to the initial

Samsung Ex. 1005, Page 378 of 1147

transmission, thus additional signaling is required so that the receiver can correctly associate each retransmission with the corresponding initial transmission. The synchronous HARQ schemes have the advantage of reducing the signaling overhead, whereas asynchronous HARQ schemes allow more flexibility in scheduling. In an adaptive HARQ scheme, transmission attributes, such as the modulation, coding, and resource allocation can be changed at each retransmission in response to variations in the channel conditions. In a non-adaptive HARQ scheme, the retransmissions are performed without explicit signaling of new transmission attributes, i.e., either the same transmission format is reused or the attributes are deterministically changed. The adaptive HARQ schemes offer higher scheduling gains at the expense of increased signaling overhead. In LTE/LTE-Advanced, asynchronous adaptive HARQ is used for the downlink and synchronous HARQ for the uplink. In the uplink, the retransmissions may be either adaptive or nonadaptive depending on the signaling of new transmission attributes.

In the multiplexing and demultiplexing entity, data from several logical channels is multiplexed or demultiplexed to/from one transport channel. The multiplexing entity generates MAC PDUs from MAC SDUs when radio resources are available for a new transmission based on the decisions of the logical channel prioritization entity. The demultiplexing entity reassembles the MAC SDUs from MAC PDUs and distributes them to the appropriate RLC entities.

The logical channel prioritization entity prioritizes the data from logical channels and decides the amount and the source of data that should be included in each MAC PDU. The decisions are sent to the multiplexing/demultiplexing entity. The random access control entity is responsible for controlling the random access procedure. The controller entity is also responsible for a number of functions including DRX, scheduling procedure, and maintaining the uplink timing alignment.

To support priority handling, multiple logical channels, where each logical channel has its own RLC entity, are multiplexed into one transport channel by the MAC sublayer. At the receiver, the MAC sublayer performs the corresponding demultiplexing and forwards the RLC SDUs to their respective RLC entity for in-sequence delivery and other RLC functions. The demultiplexing function at the receiver is enabled through addition of a MAC header, shown in Figure 8.2. There is a subheader in the MAC header associated with each RLC PDU. The subheader further contains the logical channel ID (LCID) from which the RLC PDU was originated, and the length of the PDU in octets.

In addition to multiplexing functions, the MAC sublayer can also include MAC control elements (CEs) in the transport blocks transmitted via the transport channels. A MAC control element is used for in-band control signaling, such as timing advance commands and random access response (RAR) message. The control elements are identified with reserved values in the LCID field, where the LCID value indicates the type of control information. The MAC multiplexing functions were extended to handle multiple component carriers in the case of carrier aggregation. The basic principle of carrier aggregation is independent processing of the component carriers in the physical layer, including control signaling,

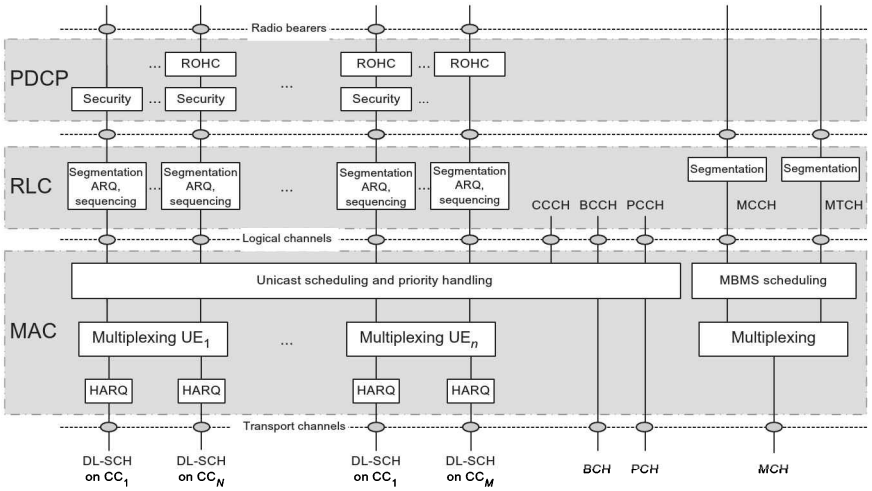


FIGURE 8.4

LTE Layer 2 structure with carrier aggregation support in the downlink [8].

scheduling, and HARQ retransmissions, while carrier aggregation is invisible to the RLC and PDCP. Carrier aggregation is therefore mainly seen in the MAC sublayer, where the logical channels, including any MAC control elements, are multiplexed to one or two transport blocks per component carrier with each component carrier having its own HARQ entity (see Figures 8.4 and 8.5).

The E-UTRA defines one MAC entity in the UE and one MAC entity in the eNB. These MAC entities control the following transport channels: BCH; DL-SCH; PCH; UL-SCH; and RACH. The exact functions performed by the MAC entities are different in the UE and the eNB. As shown in Figures 8.4 and 8.5, the MAC sublayer provides services, such as data transfer and radio resource allocation, to the upper layers. The physical layer provides the MAC sublayer with data transfer services, signaling of HARQ feedback, signaling of scheduling request, and measurements. The access to the data transfer services is through the use of transport channels. The characteristics of a transport channel are defined by its transport format which specifies the physical layer processing to be applied to the transport channel, such as channel coding and interleaving, and any other service-specific rate matching.

8.2 Logical and transport channel mapping

The LTE MAC sublayer provides data transfer services to the RLC sublayer through logical channels. Figures 8.6 and 8.7 depict classification of logical and transport channels in LTE, respectively. Each logical channel type is defined by the type of information which is transferred. The logical channels are generally classified into two groups: (1) control channels (for the transfer of control-plane

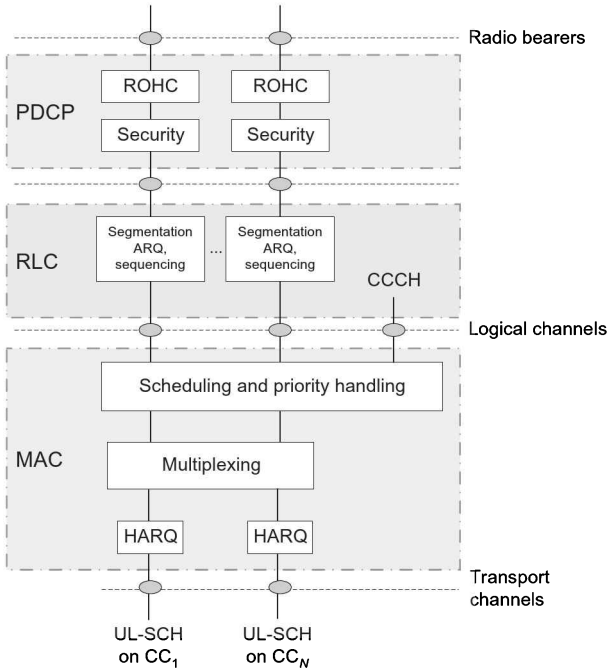


FIGURE 8.5

LTE Layer 2 structure with carrier aggregation support in the uplink [8].

information); and (2) traffic channels (for the transfer of user-plane information), as shown in Figure 8.6.

The control channels are exclusively used for transfer of control-plane information. The control channels supported by MAC can be classified as follows (Figure 8.6):

- Broadcast Control Channel (BCCH): A downlink channel for broadcasting system control information.
- Paging Control Channel (PCCH): A downlink channel that transfers paging information and system information (SI) change notifications. This channel is used for paging when the network does not know the location of the UE.
- Common Control Channel (CCCH): A logical channel for transmitting control information between UEs and eNBs. This channel is used for UEs having no RRC connection with the network.
- Multicast Control Channel (MCCH): A point-to-multipoint downlink channel used for transmitting multimedia broadcast multicast service (MBMS) control information from the network to the UE, for one or several multicast traffic channels. This channel is only used by UEs that are subscribed to MBMS.

Samsung Ex. 1005, Page 381 of 1147

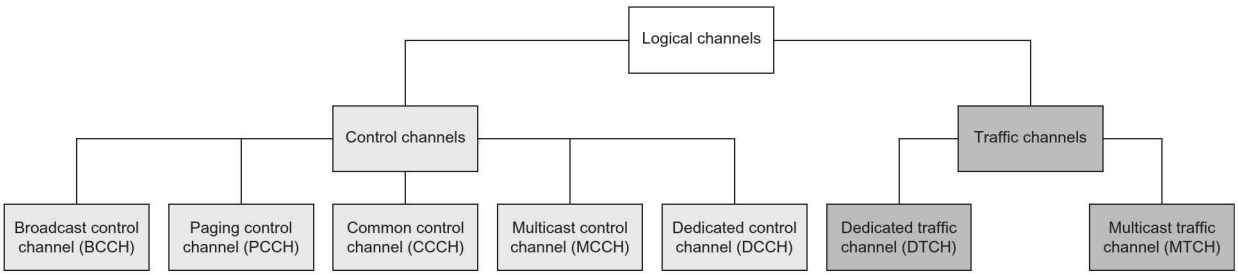


FIGURE 8.6

Classification of LTE logical channels [8].

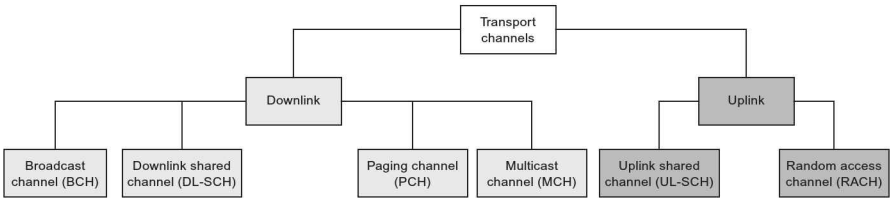


FIGURE 8.7

Classification of LTE transport channels [8].

- Dedicated Control Channel (DCCH): A point-to-point bi-directional channel that transmits dedicated control information between a UE and the network. It is used by UEs that have already established RRC connection.

The traffic channels are exclusively used for the transfer of user-plane information. The traffic channels supported by MAC can be classified as follows (as shown in Figure 8.6):

- Dedicated Traffic Channel (DTCH): A point-to-point bi-directional channel dedicated to a single UE for the transfer of user information.
- Multicast Traffic Channel (MTCH): A point-to-multipoint downlink channel for transmitting traffic data from the network to the UE. This channel is only used by UEs that receive MBMS.

As shown in Figure 8.7, downlink transport channels can be classified as follows:

- Broadcast Channel (BCH) that is characterized by a fixed, predefined transport format and is required to be broadcast in the entire coverage area of the cell.
- Downlink Shared Channel (DL-SCH) that is characterized by support for HARQ, support for dynamic link adaptation by varying the modulation, coding, and transmit power, possibility for broadcast in the entire cell, possibility to use beamforming, support for both dynamic and semi-static resource allocation, support for UE DRX to enable power saving, and support for MBMS transmission.
- Paging Channel (PCH) that is characterized by support for UE DRX in order to enable power saving, requirement for broadcast in the entire coverage area of the cell, mapped to physical resources which can also be used dynamically for traffic or other control channels.
- Multicast Channel (MCH), which is required to be broadcast in the entire coverage area of the cell, and is characterized by support for macro-diversity combining of MBMS transmission on multiple cells, and support for semi-static resource allocation.

The uplink transport channels are classified as follows (Figure 8.7):

- Uplink Shared Channel (UL-SCH), which is characterized by the possibility to use beamforming, support for dynamic link adaptation by varying the transmit

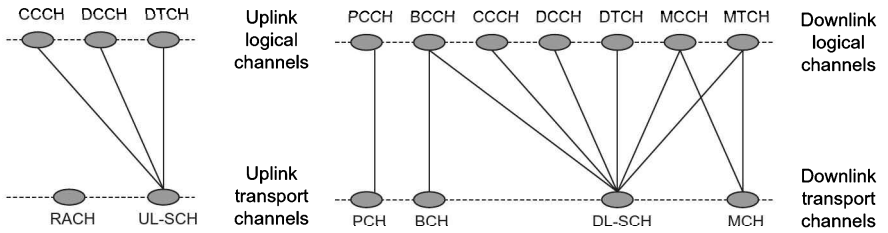


FIGURE 8.8

Mapping of logical to transport channels in the downlink and uplink [8].

power and MCS, support for HARQ, and support for both dynamic and semi-static resource allocation.

- Random Access Channel (RACH) that is characterized by limited control information and collision risk.

The mapping of logical channels to transport channels depends on the multiplexing that is configured by the RRC sublayer (see Figure 8.8) [4].

8.3 Downlink/uplink data transfer

The resource allocation on DL-SCH and UL-SCH is performed by the eNB scheduler. In the downlink, the allocation information pertaining to the physical downlink shared channel (PDSCH) within the same subframe is carried on the physical downlink control channel (PDCCH). The UE may be addressed using different radio network temporary identifiers (RNTIs). During normal operation, either cell radio network temporary identifier (C-RNTI) or semi-persistent scheduling (SPS) C-RNTI may be used depending on whether the transmission is dynamically scheduled, or is periodic with a predefined scheduling pattern. During the RACH procedure, DL-SCH transmissions may also be scheduled using the random access RNTI (RA-RNTI) or the temporary C-RNTI. Furthermore, if the system information is being transmitted via DL-SCH, then the SI-RNTI is utilized. The allocation information that is sent on the PDCCH using various downlink control information (DCI) formats includes resource block allocation, MCS, and HARQ information. Only one HARQ process may be scheduled in any subframe corresponding to either one or two transport blocks [10–12].

The allocation of the uplink grant is managed by the eNB. The allocation information is carried in the PDCCH. Similar to the downlink, the uplink transmissions can be identified using different RNTIs. For normal operation, either C-RNTI or SPS-C-RNTI may be used depending on the scheduling type. During the contention resolution (part of RACH procedure), temporary C-RNTI may be used to address a UE. The allocation information for the uplink is sent using a single DCI format that includes the allocated resource blocks, MCS, HARQ information, and whether frequency hopping is enabled.

The BCCH is a downlink logical channel that carries system information. The MAC sublayer maps the BCCH channel to BCH and DL-SCH transport channels. When mapped to BCH, BCCH broadcasts a master information block (MIB). The MIB carries essential parts of the system information that are necessary for initial system acquisition, such as system bandwidth, system frame number (SFN), and physical hybrid-ARQ indicator channel (PHICH) configuration information. The periodicity of MIB transmission is 40 ms. The remaining parts of the system information are transmitted in the form of system information blocks (SIBs), which are sent via BCCH and are mapped to DL-SCH. The SIBs are sent via system information messages, which can contain multiple SIBs. The system information messages are recognized on the DL-SCH by SI-RNTI. The system information message transmission is scheduled at regular intervals as specified in SIB1, which is transmitted on DL-SCH every 80 ms [10–12].

All SIBs are carried on the BCCH and are mapped to DL-SCH. The BCCH carries SIB1 that is mapped to the SI message 1 (SI-1), whereas other SIBs are mapped to other system information messages. The SI messages are identified by SI-RNTI, which has a predefined value of *0xFFFF*. The SIB1 is always transmitted with a period of 80 ms and contains the scheduling and multiplexing information of other SIBs. A broadcast HARQ process is utilized for all system information transmissions such that ACK/NACK feedback is not required. The redundancy version of a received downlink assignment in a specific TTI is determined by $RV_k = \lceil (3k/2) \rceil \bmod 4$ where k depends on the type of SI message. For the SIB1 message $k = (\mathrm{SFN}/2) \bmod 4$. For all other SI messages $k = l \bmod 4|_{l=0,1,\ldots,N_{\mathrm{SW}}-1}$ where l denotes the subframe number within the SI window, which has a duration defined by N_{SW}. The SI window is defined in SIB1 and specifies the starting point and time duration during which the SIs will be transmitted [11].

The paging control channel is a logical channel that is mapped by the MAC sublayer to the paging transport channel, which is transmitted on DL-SCH. For PCH, a unique radio network paging identifier with a value of *0xFFFE* is signaled in the PDCCH allocation message. Multiple UEs may be paged simultaneously using a single paging allocation with the UE identity indicated within the paging message.

8.3.1 HARQ principles

While the ARQ error control mechanism is simple and provides high transmission reliability, the throughput of ARQ schemes drops rapidly with increasing channel error rates and the ARQ latency, due to retransmissions, could be excessively high and intolerable for some delay-sensitive applications. Systems using forward error correction (FEC), on the other hand, can maintain constant throughput regardless of channel error rate. However, FEC schemes have some drawbacks. High reliability is hard to achieve with FEC and requires the use of long and powerful error correction codes that increase the complexity of implementation [10–12]. The drawbacks of ARQ and FEC can be overcome, if the two error control schemes are properly combined.

In order to achieve increased throughput and lower latency in packet transmission, the HARQ scheme was designed to combine the ARQ error control mechanism and FEC coding. A HARQ system consists of an FEC subsystem contained in an ARQ system. In this approach, the average number of retransmissions is reduced by using FEC through correction of the error patterns that occur more frequently; however, when the less frequent error patterns are detected, the receiver requests a retransmission where each retransmission carries the same or some redundant information to help the packet detection. The HARQ uses FEC to correct a subset of errors at the receiver and relies on error detection to detect the remaining errors. Most practical HARQ schemes utilize CRC codes for error detection and convolutional or turbo codes for error correction. The HARQ schemes are typically classified into three groups depending on the content of subsequent retransmissions, as follows [10–12]:

1. Type I HARQ: In type I HARQ schemes, the same data packet is transmitted in all the retransmissions. Soft combining may be used to improve the reliability. This type of HARQ is also referred to as chase combining (CC) where the blocks of data along with the CRC code are encoded using a FEC encoder before transmission. If the receiver is unable to correctly decode the data block, a retransmission is requested. When a retransmitted coded block is received, it is combined with the previously received block corresponding to the same information bits (using for example maximal ratio combining method) and fed to the decoder. Since each retransmission is an identical replica of the original transmission, the received E_b/N_0, i.e., the energy per information bit divided by the noise spectral power density, increases per each retransmission, improving the likelihood of correct decoding. In chase combining HARQ, the redundancy version of the encoded bits is not changed from one transmission to the next; therefore, the puncturing pattern remains the same. The receiver uses the current and all previous HARQ transmissions of the data block in order to decode the information bits. The process continues until either the information bits are correctly decoded and pass the CRC test, or the maximum number of HARQ retransmissions is reached. When the maximum number of retransmissions is reached, the MAC sublayer resets the process and continues with fresh transmission of the same data block.

   A number of parallel channels for HARQ can help improve the throughput as one process is awaiting an acknowlegment; another process can utilize the channel and transmit other subpackets. Figure 8.9 illustrates the operation of the chase combining HARQ scheme and how the retransmission of the same coded bits changes the combined energy per bit E_b while maintaining the effective code rate intact.
2. Type II HARQ: In the original type II HARQ scheme, an alternate parity/user data retransmission scheme is used. With proper choice of the channel encoder, the user information may also be recovered by parity bits. However, most type II algorithms adopt an incremental parity retransmission scheme, in which only additional parity bits are sent in subsequent retransmissions.

Samsung Ex. 1005, Page 386 of 1147

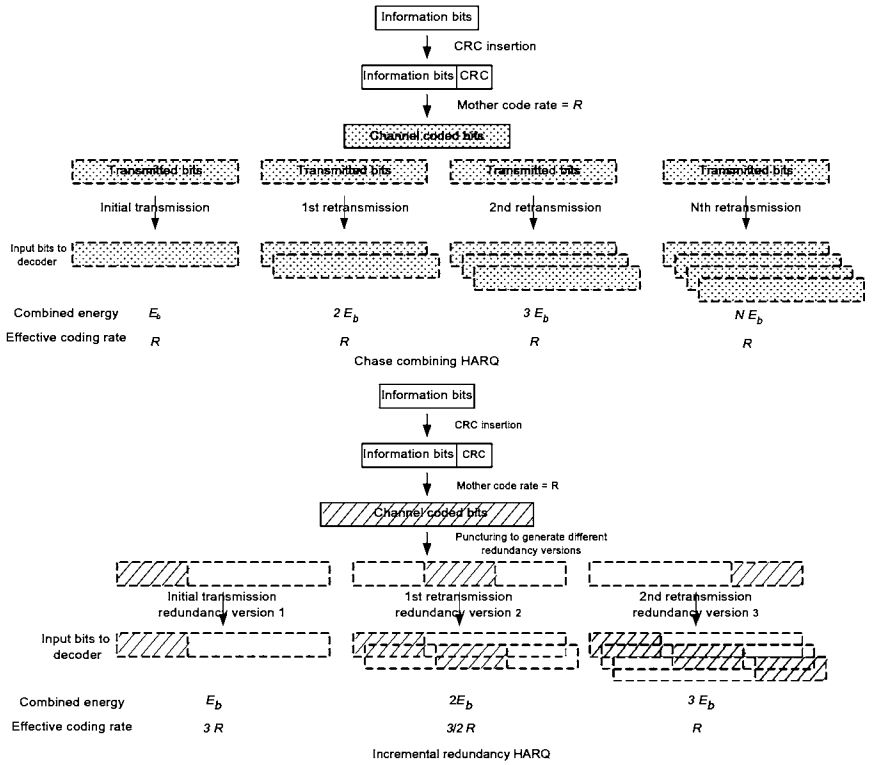


FIGURE 8.9

Illustration of the chase combining (CC) and incremental redundancy (IR) HARQ schemes [12].

Therefore, after each retransmission, a richer set of parity bits is available at the receiver, improving the probability of reliable decoding. In incremental schemes, however, information cannot be recovered from parity bits alone.

In an incremental redundancy HARQ scheme, a number of coded bits with increasing redundancy, where each represents the same set of input bits, are generated and transmitted to the receiver, when a retransmission, is requested to assist the receiver with the decoding of the information bits. The receiver combines each retransmission with the previously received bits belonging to the same packet. Since each retransmission carries additional parity bits, the effective code rate is lowered by each retransmission, as shown in Figure 8.9. The incremental redundancy is based on a low-rate code and the different redundancy versions are generated by puncturing the channel coder output. In the example shown in Figure 8.9, the basic code rate is R and 1/3 of the coded bits are transmitted in each retransmission. Aside from increasing the received signal to noise ratio E_b/N_0 by each retransmission due to combining, there is

a coding gain[1] attained as result of each retransmission. It must be noted that chase combining is a special case of incremental redundancy HARQ where the retransmissions are identical copies of the original coded bits.

3. Type III HARQ: In a type III HARQ scheme, both user data and complementary parity bits are included in every retransmission. The type III HARQ schemes are less efficient than incremental redundancy schemes due to repeated transmission of user data bits. Also in this case, after every retransmission, a richer set of parity bits is available at the receiver, improving the probability of reliable decoding. Moreover, by adopting combining techniques, reliability may be further improved. In a type III incremental scheme, information can be recovered from every transmission. In a self-decodable HARQ retransmission scheme, the user data may be recovered from every single transmission.

8.3.2 Downlink HARQ timing and signaling protocols

LTE uses an adaptive asynchronous HARQ scheme in the downlink. Therefore, the retransmissions are scheduled in the same way as new transmissions, i.e., the transmission/retransmissions may occur at any time and at an arbitrary frequency location over the downlink bandwidth. The use of asynchronous and adaptive HARQ in the LTE downlink was partially motivated by the desire to avoid scheduling conflicts with transmission of SI, multicast, and broadcast content over single frequency network (MBSFN) subframes. Instead of dropping a retransmission that otherwise would have collided with MBSFN subframes or SI, the eNB can schedule the retransmission in time and/or frequency to avoid the overlap in resources. The HARQ scheme applies to the data transmissions on PDSCH, where the ACKs are received via the physical uplink control channel (PUCCH) or physical uplink shared channel (PUSCH). There is one HARQ entity at the UE for each serving cell. There are multiple serving cells and multiple HARQ entities when the UE supports carrier aggregation [12].

Soft combining is supported through transmission of an explicit new data indicator (NDI) flag, toggled for each new transport block. In addition to the NDI, HARQ-related downlink control signaling consists of the HARQ process number (3 bits for FDD, 4 bits for TDD) and the redundancy version (2 bits), both explicitly signaled in the scheduling assignment for each downlink transmission. Downlink spatial multiplexing implies transmission of two transport blocks in parallel on a component carrier. In this case, it is possible to retransmit only one of the transport blocks, since error occurrences in transmission of the transport blocks are typically uncorrelated; thus each transport block has separate new data and redundancy version indicators. However, there is no need to signal the process number separately, given that each HARQ process consists of two sub-processes in

[1]The coding gain is defined as the difference between E_b/N_0 required to achieve a given bit error rate (BER) in a coded system and the E_b/N_0 required to achieve the same BER in an uncoded system.

the case of spatial multiplexing, once the process number for the first transport block is known, the process number for the second transport block is derived implicitly. Transmissions on downlink component carriers are acknowledged independently. On each component carrier, transport blocks are acknowledged by transmitting one or two bits on the uplink. In the absence of spatial multiplexing, there is only a single transport block within a TTI and consequently only a single acknowledgment bit is required. However, in the case of spatial multiplexing, there are two transport blocks per TTI, each requiring its own HARQ feedback. The total number of bits required for HARQ acknowledgments depends on the number of component carriers and the transmission mode of each component carrier. Since downlink component carriers are scheduled separately using the PDCCH, the HARQ process numbers are signaled independently for each component carrier. It is obvious that the terminal is not expected to transmit a HARQ acknowledgement in response to transmission of system information, paging messages, and other broadcast information. Therefore, HARQ feedback is only sent in the uplink for unicast transmission [12].

LTE utilizes incremental-redundancy HARQ with a 1/3 turbo encoder used for FEC. The transport block CRC is used to detect errors. The receiver only receives different punctured versions of the same turbo-encoded data; each of these retransmissions is self-decodable. Thus, it falls into the category of a type III HARQ described earlier. The maximum number of retransmissions is limited to four, with an initial coding rate of 1/2 or 3/4. The maximum number of simultaneous downlink HARQ processes (number of PDSCH transmissions) is limited to eight for FDD mode, as specified in [7].

8.3.3 Uplink HARQ timing and signaling protocols

The operation of the LTE uplink HARQ protocol is different from the downlink counterpart. The choice of uplink HARQ protocol was made by taking into consideration the lower overhead of a synchronous non-adaptive scheme compared to an asynchronous adaptive scheme. As a result, uplink retransmissions always occur at fixed intervals. In the case of FDD mode, the uplink retransmissions occur eight subframes after the previous transmission attempt for the same HARQ process. The set of resource blocks used for the retransmission on each component carrier are identical to the initial transmission. The control signaling required in the downlink for an uplink retransmission is a HARQ feedback, which is transmitted on PHICH. In the case of a NACK on PHICH, the data is retransmitted. Spatial multiplexing is handled in the same way as in the downlink where two transport blocks are concurrently transmitted in the uplink, each with its own MCS and NDI, but sharing the same HARQ process number. The two transport blocks are individually acknowledged, thus two bits of information are needed in the downlink to acknowledge an uplink transmission when spatial multiplexing is utilized. The uplink carrier aggregation is another example where multiple ACKs/NACKS are required, i.e., one or two ACK/NACK bits per uplink component

carrier. In the TDD mode, the uplink transmissions in different subframes may need to be acknowledged in the same downlink subframe. Thus, multiple PHICHs may be needed as each PHICH is capable of transmitting one bit of HARQ feedback. Each of the PHICHs is transmitted on the same downlink component carrier that was used to schedule the initial uplink transmission. Despite the fact that the default mode of operation for the uplink is synchronous, non-adaptive HARQ, it is also possible to operate the uplink HARQ in a synchronous, adaptive mode, where the resource block set and MCS for the retransmissions are changed according to the transmission conditions. Although non-adaptive retransmissions are typically used, due to their very low downlink control signaling overhead, adaptive retransmissions are sometimes useful in order to avoid the uplink frequency resource fragmentation or to avoid collisions with random access allocated resources. A UE is scheduled for an initial transmission in subframe n and if the transmission is not correctly received, a retransmission is required in subframe $n + 8$ in FDD mode (the TDD retransmission timing depends on the subframe configuration). With non-adaptive HARQ, the retransmissions occupy the same part of the uplink spectrum as the initial transmission. In this case, the spectrum may be fragmented, which limits the bandwidth available for other terminals, unless the other terminal is capable of multi-cluster transmission (LTE Rel-10 and onward UEs). Example HARQ operation and timing in the uplink for an FDD and TDD system are shown in Figures 8.10 and 8.11, respectively.

Figure 8.12 depicts an example uplink HARQ process using adaptive and non-adaptive schemes to better understand this operation. In the adaptive case, the

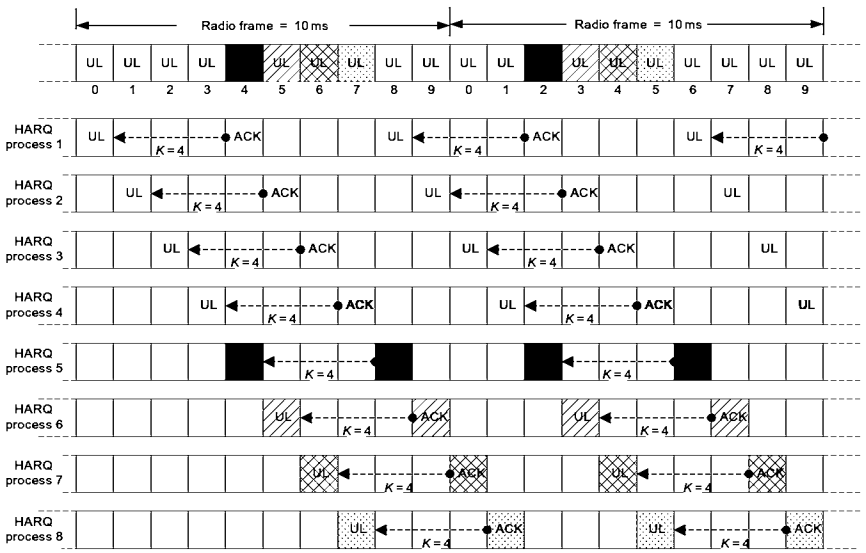


FIGURE 8.10

Example uplink HARQ timing for an FDD system [10].

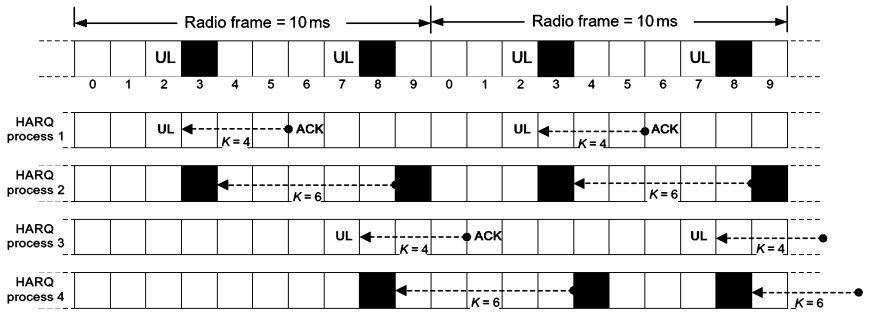


FIGURE 8.11

Example uplink HARQ timing for a TDD system with subframe configuration 1 [10].

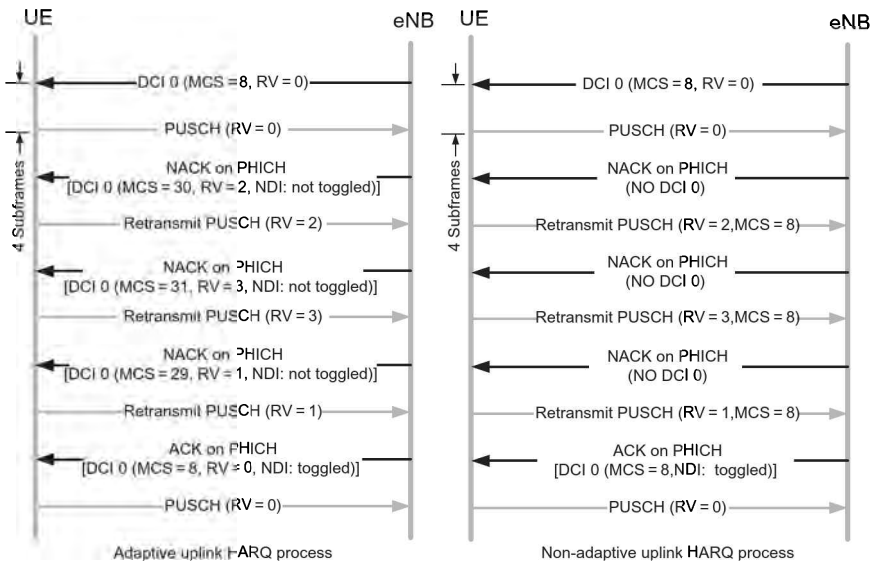


FIGURE 8.12

Illustration of example adaptive and non-adaptive uplink HARQ processes [21].

main idea is that each uplink retransmission uses a different redundancy version (RV) and the RV is determined by DCI format 0 (i.e., uplink grant). In the case of non-adaptive uplink HARQ, each uplink retransmission uses a different RV and the RV is determined by a predefined sequence specified in [1]. An important question is how the UE determines whether it is supposed to use adaptive retransmission or non-adaptive retransmission. The UE uses adaptive retransmission if it detects DCI format 0 and NDI is not toggled. In this case, the UE ignores HARQ feedback in PHICH and it retransmits based on DCI format 0 information. The

UE determines that it is supposed to use non-adaptive retransmission when it receives negative HARQ feedback in PHICH and does not receive DCI format 0.

The adaptive and non-adaptive HARQ schemes can be supported by having the terminal save the content of its transmission buffer until it receives an NDI on PDCCH. The decision on whether the data in the transmission buffer should be retransmitted is made based on the NDI included in the uplink scheduling grant which is sent in the PDCCH. The NDI is toggled for each new transmission. If the NDI is toggled, the terminal flushes the transmission buffer and transmits a new data packet. However, if the NDI does not indicate transmission of a new transport block, the previous transport block is retransmitted. The negative HARQ ACK on PHICH could instead be seen as a single-bit scheduling grant for retransmissions, where the set of bits to transmit and all the resource information are known from the previous transmission attempt [12].

The NDI is explicitly transmitted in the uplink grant. However, unlike the downlink case, the redundancy version is not explicitly signaled for each retransmission and cannot be inferred from a single-bit HARQ feedback on PHICH. Instead, given that the uplink HARQ protocol is synchronous, the redundancy version follows a predefined pattern, starting with zero when the initial transmission is scheduled by PDCCH. Whenever a retransmission is requested by sending a NACK on PHICH, the next redundancy version in the sequence is used. However, if a retransmission is explicitly scheduled by PDCCH overriding PHICH, there is a possibility that a different redundancy version may be used. Uplink grants for retransmissions use the same format as normal grants (for initial transmissions). The transport block size is already known from the initial transmission and cannot change between retransmission attempts.

The uplink HARQ retransmissions are associated with the HARQ feedback as seen by the UE on PHICH and the uplink grant sent in PDCCH. The information in the uplink grant takes precedence over retransmissions. The UE adheres to the following rules:

1. If the HARQ feedback is an ACK and no grant is detected for the HARQ process on PDCCH, then the UE maintains the data in its soft buffer until the PDCCH grant indicates whether the previous transmission was successful.
2. If the HARQ feedback is a NACK and no PDCCH grant is detected for the HARQ process, then the UE retransmits using the same PUSCH resources as the last transmission. The redundancy version used follows a known sequence.
3. If the PDCCH indicates a new transmission on PDCCH for a specific HARQ process, then regardless of HARQ feedback on PHICH, it is treated as a new transmission. Any data in the soft buffer is flushed.
4. If the PDCCH indicates a retransmission for a specific HARQ process, then irrespective of HARQ feedback, the UE performs a retransmission according to the parameters on the PDCCH.

In the case of carrier aggregation, if cross-carrier scheduling is not used and each uplink component carrier is scheduled on its corresponding downlink component carrier, the uplink component carriers are not associated with the same

PHICH resource. However, when cross-carrier scheduling is utilized, transmissions on multiple uplink component carriers may need to be acknowledged on a single downlink component carrier. In order to avoid PHICH collisions, the scheduler ensures that different reference signal phase rotations (cyclic shifts) or different resource block starting positions are used for different uplink component carriers. For semi-persistent scheduling, the reference signal phase rotation is always set to zero, since the semi-persistent scheduling is only supported on the primary component carrier. If multiple component carriers are scheduled for uplink data transmission in a subframe, there will be one or two (with spatial multiplexing) PHICHs transmitted per uplink component carrier. In general, LTE transmits the PHICH on the same component carrier that was used for the uplink grant, scheduling the corresponding uplink data transmission. Thus, a terminal is only required to monitor the component carriers that include uplink scheduling grants (especially as the PDCCH may override PHICH to support adaptive retransmissions in some cases), which helps to reduce the terminal's power consumption [10–12].

The transmission of HARQ feedback in the uplink in response to PDSCH transmissions is illustrated in Figure 8.13 for an FDD system. The data on PDSCH is transmitted to the terminal in subframe n and received by the terminal after the propagation delay in subframe n. The terminal attempts to decode the received signal, using soft combining of previous retransmissions, and then it transmits a HARQ ACK/NACK in uplink subframe $n + 4$. It must be noted that the start of an uplink subframe at the terminal is offset by the timing advance relative to the start of the corresponding downlink subframe at the terminal. Upon reception of the HARQ feedback, the eNB can transmit new data or retransmit the previous data in subframe $n + 8$. Thus, the HARQ round-trip time (RTT) is 8 ms for an FDD system.

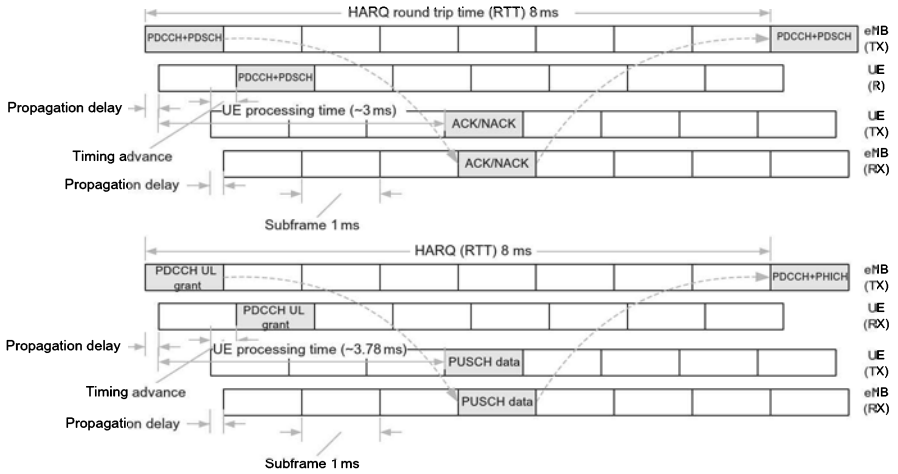


FIGURE 8.13

Example timing diagram for downlink and uplink HARQ schemes (FDD mode) [11].

In practice, the UE processing time depends on the value of the timing advance (to compensate variable propagation delay), or alternatively on its distance to the base station. Since the UE is required to operate at any distance up to the maximum cell radius of 100 km (corresponding to 0.67 ms) supported by the standard, the UE is designed for the worst-case scenario. That means UE baseband processing time is approximately 2.3 ms. For the eNB, the baseband processing time is 3 ms [11].

In LTE TDD, the time association between the data transmission and the ACK/NACK cannot be maintained due to the variable numbers of downlink and uplink subframes being present in a frame. The uplink and downlink delay between data and ACK/NACK is dependent on the TDD subframe configuration. Therefore, a fixed delay between a transmission and the HARQ ACK/NACK is not possible in LTE TDD. In TDD mode, the delay between the transmission and the HARQ ACK/NACK depends on both TDD subframe configuration and the subframe in which the data was transmitted. A fixed delay cannot be assured as subframes are allocated to downlink and uplink depending on the subframe configuration. As an example, in subframe configuration 1 (see Table 8.1), there are a number of downlink subframes whose nearest uplink subframe (four or more subframes apart) is seven subframes apart.

Table 8.1 shows the HARQ ACK/NACK timing delays for the downlink and uplink of a TDD system assuming the initial data transmission occurred in subframe n, i.e., the transmission of ACK/NACK occurs in subframe $n + k$. It can be noticed that the delays are dependent on the configuration and the subframe in which the original data transmission has occurred. For example, in configuration 1, the delay from the eNB side can range from four to six subframes, while the delays from the UE side can be from four to seven subframes. As a result, the delay for transmission of ACK/NACK can range from four to seven subframes for the downlink HARQ and from four to thirteen subframes for the uplink HARQ. The higher number in the uplink case is due to the larger downlink asymmetry in the configuration. Therefore, in TDD systems, the acknowledgment of a transport block in subframe n is transmitted in subframe $n + k$, where $k \geq 4$ and is selected such that $n + k$ is an uplink subframe when the acknowledgment is transmitted from the terminal on PUCCH or PUSCH, and a downlink subframe when the acknowledgment is transmitted from the eNB on PHICH. The value of k depends on the downlink–uplink subframe configuration, as shown in Table 8.1 for downlink and uplink transmissions.

In FDD systems, multiple HARQ processes can be activated in both downlink and uplink directions. Due to the fixed timing associations between the original and the retransmission on the uplink, the maximum number of HARQ processes has been set to eight. However, in TDD systems, due to the variable delays experienced with respect to ACK/NACK arrivals as well as the retransmissions, the number of HARQ processes is dependent on the configuration. The maximum number of HARQ processes allowed per configuration for downlink and uplink is given in Table 8.1. Typically, the retransmissions can be scheduled four subframes (or more) after the reception of a NACK message. In LTE TDD, the numbers of downlink and uplink subframes in a frame are not equal. For instance, configurations 0–5 have more downlink subframes compared to uplink subframes. Consequently, data

ble 8.1 Number of HARQ Processes and Timing for Different TDD Configurations [12]

Subframe onfiguration (DL:UL)	Downlink Subframe Number										Uplink Subframe Number										Number of Processes	
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	Uplink	Downlink
(2:3)	4	6	–	–	–	4	6	–	–	–	–	–	4	7	6	–	–	4	7	6	7	4
(3:2)	7	6	–	–	4	7	6	–	–	4	–	–	4	6	–	–	–	4	6	–	4	7
(4:1)	7	6	–	4	8	7	6	–	4	8	–	–	6	–	–	–	–	6	–	–	2	10
(7:3)	4	11	–	–	–	7	6	6	5	5	–	–	6	6	6	–	–	–	–	–	3	9
(8:2)	12	11	–	–	8	7	7	6	5	4	–	–	6	6	–	–	–	–	–	–	2	12
(9:1)	12	11	–	9	8	7	6	5	4	13	–	–	6	–	–	–	–	–	–	–	1	15
(5:5)	7	7	–	–	–	7	7	–	–	5	–	–	4	6	6	–	–	4	7	–	6	6

transmission from multiple downlink subframes needs to be acknowledged in a certain uplink subframe and vice versa for certain cases for the uplink data transmission.

The transmission of multiple ACK/NACK messages in uplink or downlink subframes is a unique feature of LTE TDD compared with FDD due to the above-mentioned limitation. Two methods have been specified in the standard to deal with this issue: ACK/NACK bundling and ACK/NACK multiplexing. For the eNB, transmission of separate ACK/NACKs is not a major problem as the eNB is not power constrained and can send multiple ACK/NACKs in a subframe. This method, where separate ACK/NACKs are sent in the same subframe for transmissions in different subframes, is called ACK/NACK multiplexing. Thus, in the downlink ACK/NACK transmission, it is by default ACK/NACK multiplexing. However, in the uplink, the UEs might be power constrained depending on their location within the cell. For UEs which are not constrained by power and coverage, multiple PUCCH messages relating to each HARQ process can be generated, i.e., separate ACK/NACK messages can be sent. For coverage-limited UEs, ACK/NACK bundling is used since fewer bits need to be sent in this case. Multiplexing implies that independent ACKs/NACKs for each of the received transport blocks are sent to the eNB, which allows independent retransmission of erroneous transport blocks. However, it also suggests that multiple bits need to be transmitted from the terminal, which may limit the uplink coverage. This is the motivation for the bundling mechanism.

Bundling of ACKs/NACKs implies that the outcome of the decoding of downlink transport blocks from multiple downlink subframes can be combined into a single HARQ feedback transmitted in the uplink [12]. A logical AND operation on all ACK/NACK messages is typically performed and the result is sent. For example, in TDD subframe configuration 1, the transmissions in subframes 0 and 1 are acknowledged in subframe 7. A logical AND of these messages is performed and sent. This implies that both transport blocks will be retransmitted by the eNB as it does not have the individual acknowledgment status of the downlink transmissions. This is a disadvantage of ACK/NACK bundling as more retransmissions than necessary might be sent. The main advantage of the ACK/NACK bundling method is the fact that the limited power resources of a UE will now be used only in a fewer resource blocks as compared to a larger number of resource blocks. This can improve the uplink coverage of the UE since there will be more power per resource block that will be available to the UE in the case of ACK/NACK bundling [12].

In order to transmit on the UL-SCH the UE must have a valid uplink grant (except for non-adaptive HARQ retransmissions) which it may receive dynamically on the PDCCH or in an RAR message. To perform requested transmissions, the MAC sublayer receives HARQ information from lower layers. When the physical layer is configured for uplink spatial multiplexing, the MAC sublayer can receive up to two grants (one per HARQ process) for the same TTI from lower layers [1].

At a given TTI, if an uplink grant is indicated for the TTI, the HARQ entity identifies the HARQ processes for which a transmission should occur. It also routes the received HARQ feedback (ACK/NACK information), modulation and

Samsung Ex. 1005, Page 396 of 1147

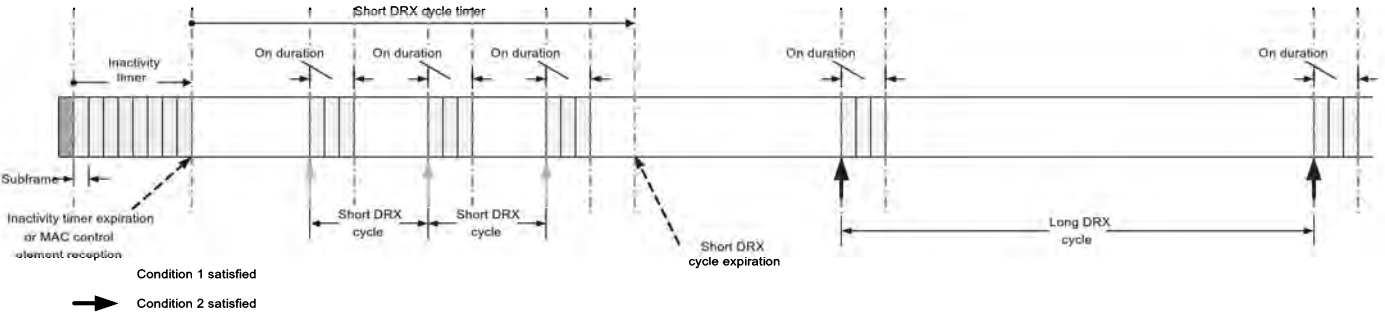
coding, and resource allocation information (conveyed by the physical layer) to the appropriate HARQ processes. When TTI bundling is configured, the parameter *TTI_BUNDLE_SIZE* provides the number of TTIs of a TTI bundle. The TTI bundling operation relies on the HARQ entity for invoking the same HARQ process for each transmission that is part of the same bundle. Within a bundle, HARQ retransmissions are non-adaptive and triggered without waiting for feedback from previous transmissions according to the *TTI_BUNDLE_SIZE*. The HARQ feedback of a bundle is only received for the last TTI of the bundle (i.e., the TTI corresponding to *TTI_BUNDLE_SIZE*), regardless of a transmission in that TTI (e.g., when a measurement gap occurs). A retransmission of a TTI bundle is also a TTI bundle. TTI bundling is not supported when the UE is configured with one or more secondary cells (SCells) with configured uplink [1].

Each HARQ process in the uplink maintains a state variable *CURRENT_TX_NB*, which indicates the number of transmissions that have taken place for the MAC PDU currently in the buffer, and a state variable *HARQ_FEEDBACK*, which indicates the HARQ feedback for the MAC PDU currently in the buffer. When the HARQ process is established, *CURRENT_TX_NB* set to zero. The sequence of redundancy versions is 0, 2, 3, and 1 (Figure 8.12). The variable *CURRENT_IRV* is an index to the sequence of redundancy versions. This variable is updated in modulo 4. New transmissions are performed on the resource and with the MCS indicated on PDCCH or RAR message. Adaptive retransmissions are performed on the resource and, if provided, with the MCS indicated on PDCCH. Non-adaptive retransmission is performed on the same resource and with the same MCS as was used for the last transmission attempt [1].

8.4 DRX operation

LTE supports DRX in the downlink, to save the UE battery power during periods of inactivity. In the DRX mode, the UE only monitors PDCCH during a predefined on-duration interval within one of two possible DRX cycles. The UE may turn off its receiver when it does not expect to receive any transmission in the downlink over PDCCH. The UE can transition to DRX mode in one of the following two ways: (1) implicitly based on the expiration of certain configured timers; or (2) explicitly based on the transmission of a DRX command in the form of a MAC control element.

As shown in Figure 8.14, there are a number of timers associated with the DRX mode. The DRX inactivity timer controls the number of subframes for which a UE continues to monitor the downlink before an absence of activity triggers the UE to enter the DRX mode. Upon entering the DRX mode, the UE monitors the downlink for a number of subframes defined by the on-duration timer in every short DRX cycle or long DRX cycle. The short DRX cycle is optional. If it is configured, the UE will follow the short DRX cycle for a specific number of cycles controlled by the short DRX cycle timer before switching to the long DRX



GURE 8.14

ustration of short and long DRX cycles and timers [10].

cycle, if there is no activity. Other timers prevent the UE from entering the DRX mode. These are associated with HARQ downlink retransmissions. If the UE has received data that it was unable to decode and this was not associated with a broadcast HARQ process, then following the expiration of the HARQ RTT timer, which specifies the earliest time at which a UE can expect a retransmission, the DRX retransmission timer is started. The UE must wait for this timer to expire and continue to monitor the downlink before it can enter the DRX mode. Moreover, the UE cannot enter the DRX mode under the following conditions:

- If downlink transmissions are in progress or if the UE has sent a scheduling request.
- If the UE is expecting an uplink grant for HARQ retransmission.
- If the RACH procedure is in progress and the contention resolution timer is running.

In addition to the above timers, the network can order the UE to enter the DRX mode with the DRX command MAC control element. All timers and parameters described in this context are configured by RRC signaling.

A UE may be configured by the RRC sublayer for DRX, which controls the way the UE monitors PDCCH and looks for its C-RNTI, TPC-PUCCH-RNTI, TPC-PUSCH-RNTI, and SPS-C-RNTI (if configured). When the UE is in the RRC_CONNECTED state and if DRX is configured, the UE is allowed to monitor PDCCH intermittently; otherwise, the UE monitors PDCCH continuously. The RRC signaling controls DRX operation by configuring the following timers: *onDurationTimer*; *drx-InactivityTimer*; *drx-RetransmissionTimer* (for each downlink HARQ process except for the broadcast process); *longDRX-Cycle*; *drxStartOffset*; and optionally *drxShortCycleTimer* and *shortDRX-Cycle*. Note that the same activity time applies to all activated serving cells in carrier aggregation scenarios [1].

In Figure 8.14, Conditions 1 and 2 are defined as follows:

Condition 1: $(10\text{SFN} + n_{\text{subframe}})\text{mod}(shortDRX\text{-}Cycle) = (drxStartOffset) \text{mod}(shortDRX\text{-}Cycle)$.

Condition 2: $(10\text{SFN} + n_{\text{subframe}})\text{mod}(longDRX\text{-}Cycle) = drxStartOffset$.

The DRX parameters and timers in the above equations are defined as follows [1]:

- *Active time*: The time interval during which the UE monitors the PDCCH in subframes containing a PDCCH.
- *DRX cycle*: The periodic repetition of the on-duration followed by a potential period of inactivity.
- *drx-InactivityTimer*: The number of consecutive subframes containing a PDCCH, after the subframe in which a PDCCH indicates an initial uplink/downlink user data transmission for the UE.
- *drx-RetransmissionTimer*: The maximum number of consecutive subframes containing a PDCCH immediately after a downlink retransmission is expected by the UE.

- *drxShortCycleTimer*: The number of consecutive subframes that the UE is required to follow the *short DRX cycle*.
- *drxStartOffset:* Specifies the subframe where the DRX cycle starts.
- *onDurationTimer*: The number of consecutive subframes containing a PDCCH at the beginning of a DRX cycle. This is the time duration after waking up from DRX that the UE monitors the PDCCH region in the subframe. If the UE successfully decodes a PDCCH, the UE stays awake and starts the inactivity timer.

From the above parameters, the on-duration and inactivity timer have fixed lengths, while the active time has variable length based on scheduling decision and the UE decoding success. Only on-duration and inactivity timer duration are signaled to the UE by the eNB. There is only one DRX configuration applied in the UE at any time and the UE is required to apply an on-duration upon wake-up from the DRX cycle. This is also applicable for the case where the UE has only one real-time service that is being handled through the allocation of predefined resources. This would allow other signaling, mechanisms, such as RRC signaling, to be sent during the remaining portion of the active time. The new transmissions can only occur during the active time, so that when the UE is waiting for one retransmission, it does not have to be awake during the HARQ round-trip time. If the PDCCH has not been successfully decoded during the on-duration, the UE must follow the DRX configuration, i.e., the UE can enter the DRX cycle, if allowed by the DRX configuration. This applies also for the subframes where the UE has been allocated predefined resources.

If the UE successfully decodes a PDCCH for a first transmission, it stays awake and starts the inactivity timer, even if a PDCCH is successfully decoded in the subframes where the UE has also been allocated predefined resources, until a MAC control message informs the UE to reenter the DRX cycle, or until the inactivity timer expires. In both cases, if a short DRX cycle is configured, the UE first follows the short DRX cycle and after a longer period of inactivity the UE follows the long DRX cycle; otherwise, the UE directly enters the long DRX cycle. When the DRX is configured, the network should detect whether the UE remains in the E-UTRAN coverage by requesting the UE to send periodic signals to the network. In the case of carrier aggregation, whenever a UE is configured with only one serving cell, i.e., primary cell, LTE Rel-8/9 DRX scheme applies. In other cases, the same DRX operation applies to all configured and activated serving cells (i.e., identical active time for PDCCH monitoring) [8].

During the DRX procedure, the UE maintains a DRX cycle that is defined as a number of subframes. The UE monitors the PDCCH for a specific number of subframes in a DRX cycle. This time interval is called on-duration and can range from 1 to 200 subframes. The UE may turn off its receiver for rest of the DRX cycle. The UE maintains two DRX cycles, referred to as short DRX cycle and long DRX cycle, which have different durations. The short DRX cycle is optional and targeted for applications, such as voice over IP (VoIP) which typically require transmission of relatively small amounts of data at short but regular intervals. If configured, the UE starts with a short DRX cycle (2−640 subframes) when it enters the DRX mode.

When the configurable short DRX timer expires, the UE transitions to the long DRX cycle (10–2560 subframes). The UE can transition to the DRX mode either upon expiration of configured timers or following a DRX command in the form of a MAC control element. Figure 8.15 illustrates different scenarios in DRX operation.

The downlink radio link quality of the primary cell must be monitored by the UE for the purpose of indicating out-of-sync/in-sync status to higher layers. In non-DRX operation mode, the physical layer in the UE in every radio frame assesses the radio link quality over the previous time period against thresholds (Q_{out} and Q_{in}). In DRX operation mode, the physical layer in the UE, at least once every DRX period assesses the radio link quality over the previous time period, against thresholds (Q_{out} and Q_{in}). The threshold Q_{out} is defined as the level at which the downlink radio link cannot be reliably received, corresponding to a 10% block error rate of a hypothetical PDCCH transmission taking into account the

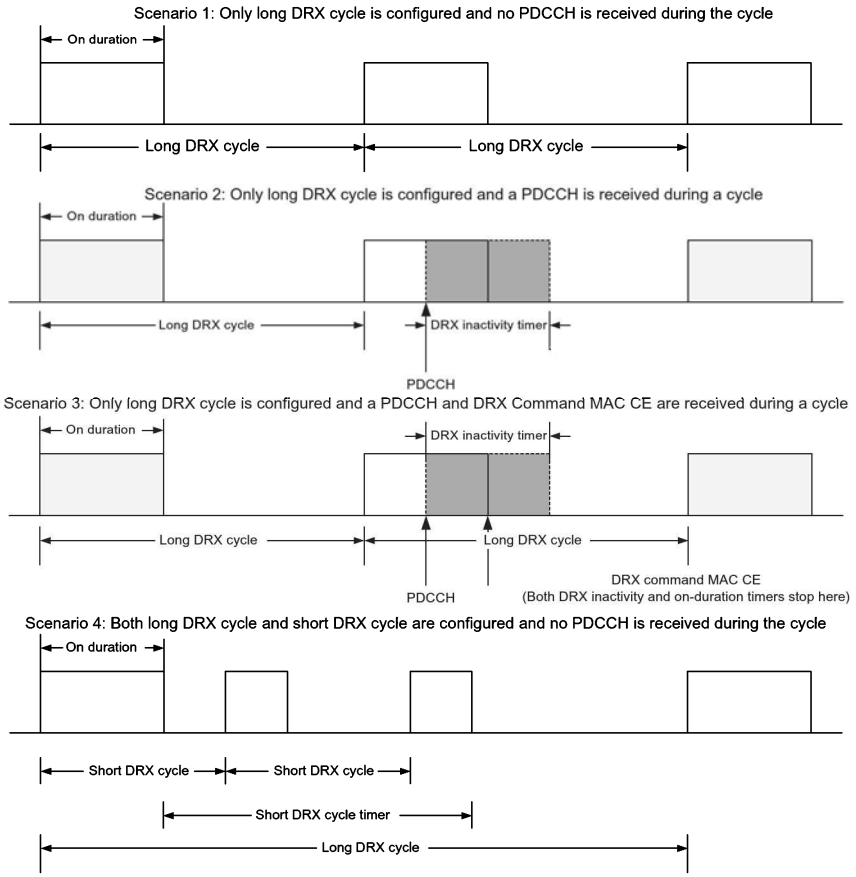


FIGURE 8.15

Illustration of DRX operation [21].

physical control format indicator channel (PCFICH) errors with transmission parameters specified in [9]. The threshold Q_{in} is defined as the level at which the downlink radio link quality can be significantly more reliably received than at Q_{out} and corresponds to a 2% block error rate of a hypothetical PDCCH transmission taking into account PCFICH errors with transmission parameters specified in [9].

To investigate the effect of DRX on packet delay, let's assume that the inter-packet arrival times $T_{arrival}$ follow an exponential distribution with a mean of $1/\lambda$ ms, the extra delay T_d (in milliseconds), caused by active mode of DRX, can be computed as $T_d = T_{arrival} \text{mod} T_{DRX\text{-}cycle}$. The probability distribution of the extra delay T_d can be expressed as follows [13]:

$$p_d(k) = \frac{1}{T_{DRX\text{-}cycle}} \sum_{i=0}^{T_{DRX\text{-}cycle}-1} \frac{\lambda}{a_i^k(1 - a_i e^{-\lambda})} \tag{8.1}$$

where a_i is the ith order-n root of unity[2] and $p_d(k)$ is the probability that the extra delay is equal to k subframes.

The energy saving of the UE due to use of the DRX mechanism can be expressed as follows [13]:

$$Energy\_Saving\% = 100 \frac{N_{DRX}E_{DRX} + N_{Active}E_{Active}}{(N_{DRX} + N_{Active})E_{Active}} \tag{8.2}$$

It is assumed that there are N_{DRX} frames during which the UE is in the DRX mode and N_{Active} frames during which the UE is in the normal operation mode. Furthermore, we assume that the energy consumed per frame is E_{DRX} and E_{Active} during the DRX and active modes, respectively. The ratio of E_{DRX} and E_{Active} is directly related to the number of hardware sub-systems (or modules), that are powered down during the DRX mode [13,14,18].

8.5 MAC control elements

MAC control element are control commands and reports that enable MAC operation. They are transmitted as part of the DL-SCH or UL-SCH and can be piggybacked to the data payloads. The following CEs have been defined in MAC specification as part of LTE Rel-11:

Downlink

- Timing advance: This CE is used to provide initial and periodic timing adjustment commands to the UE in the uplink.

[2]In mathematics, a root of unity, often called a de Moivre number, is a complex number whose nth power is equal to one. Roots of unity are used in many branches of mathematics, and are especially important in number theory, field theory, and discrete Fourier transform. The notion of root of unity also applies to any algebraic ring with a multiplicative identity element, namely a root of unity is any element of finite multiplicative order. An nth root of unity, where $n = 1, 2, 3, \ldots$ is a positive integer, is a complex number z satisfying the equation $z^n = 1$. An nth root of unity is primitive, if it is not a kth root of unity for some smaller integer k, i.e., $z^k \neq 1 (k = 1, 2, 3, \ldots, n-1)$ [20].

Samsung Ex. 1005, Page 402 of 1147

- DRX command: This CE initiates DRX mode at the UE and conveys DRX commands to the UE.
- UE contention resolution identity: This CE is used during the RACH procedure to resolve possible contention between multiple UEs that are trying to simultaneously access the network.
- MBMS dynamic scheduling information: This CE is transmitted for each MCH to inform the MBMS-capable UEs of scheduled data transmission on MTCH.
- Activation/deactivation: This CE is used to activate or deactivate secondary cells for the carrier-aggregation-enabled UEs.

Uplink

- Buffer status report (BSR): This CE is used to report the UE's soft buffer status and to enable uplink scheduling.
- Power headroom: This CE is used to report the UE's transmit power relative to maximum permissible value, or whether the UE is currently power limited.
- C-RNTI report: This is used to identify a UE when sending information over CCCH and for the UE to transmit a dedicated C-RNTI during the random access procedure for the purpose of contention resolution.
- Extended power headroom: This CE was specified to support carrier aggregation in LTE Rel-10. In the case of carrier aggregation, the UE reports power headroom of each component carrier independently.

As shown in Figure 8.16, each MAC CE included in the MAC PDU has an associated one-byte subheader in the MAC header with the format R/R/E/LCID, where the LCID indicates the type of MAC CE. The values of LCID for DL-SCH and UL-SCH are listed in Table 8.2.

8.5.1 Uplink timing advance control element

After initial synchronization on RACH, the UE must maintain a reasonably accurate uplink time alignment. This is achieved using the timing advance control element which is sent as part of a MAC PDU on DL-SCH, enabling the eNB to periodically send timing adjustment commands to the UE. The network may also send timing adjustment commands as part of an RAR message. When the

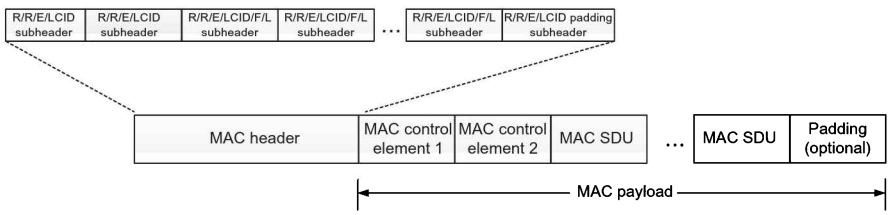


FIGURE 8.16

An example MAC PDU containing MAC header, MAC CEs, MAC SDUs, and padding [1].

Table 8.2 Values of the LCID for DL-SCH and UL-SCH [1]

Values of LCID for DL-SCH		Values of LCID for UL-SCH	
Index	**LCID Values**	**Index**	**LCID Values**
00000	CCCH	00000	CCCH
00001–01010	Identity of the logical channel	00001–01010	Identity of the logical channel
01011–11010	Reserved	01011–11000	Reserved
11011	Activation/deactivation	11001	Extended PHR
11100	UE contention resolution identity	11010	PHR
11101	Timing advance command	11011	C-RNTI
11110	DRX command	11100	Truncated BSR
11111	Padding	11101	Short BSR

command is sent as a control element, the timing advance is a 6-bit number that adjusts timing (advances or delays) relative to the current UE radio frame timing. If the command is sent via an RAR message, the timing advance is an 11-bit number that specifies absolute timing relative to the downlink radio frame. The eNB calculates the preferred timing advance based on uplink transmissions from the UE. The eNB also configures the UE's timing alignment timer. If the UE does not receive a timing advance MAC control element before the timing alignment timer expires, it is no longer considered to be uplink-synchronized with the cell.

The UE must apply the timing advance command when it receives it from the eNB. The command received in the nth frame is applied in the $(n + 6)$th frame. The command provides incremental value which adjusts the current timing alignment value. The incremental value may be negative or positive. Positive value advances and negative value delays the uplink transmission timing. If the UE does not receive a timing advance MAC control element before the timing alignment timer expires, it is no longer considered to be synchronized with the cell. When the timer expires, the UE must flush HARQ buffers and reinitialize the HARQ processes, release all PUCCH and sounding reference signal resources. If any uplink transmission is required, the UE must perform contention-based RACH synchronization procedures to obtain timing alignment and an uplink grant before it can transmit in the uplink.

The timing advance command MAC control element in LTE Rel-11 has a fixed size and consists of a single octet as shown in Figure 8.17. The information fields in this control element consist of a timing advance group[3] (TAG) identity (TAG ID), which indicates the TAG identity of the addressed TAG and the timing advance command. The TAG containing the PCell has the TAG identity value of zero. The

[3]A timing advance group is a group of serving cells that is configured by RRC signaling and uses the same timing reference cell and the same timing advance value. Primary timing advance group (pTAG) is the timing advance group containing the PCell. Secondary timing advance group (sTAG) is the timing advance group not containing the PCell [8].

length of this field is 2 bits. The timing advance command indicates the index value of timing advance $T_A = (0, 1, 2, \ldots, 63)$ that provides the amount of timing adjustment that the UE has to apply. The length of this field is 6 bits. Note that in the earlier releases of LTE, the leftmost 2 bits were reserved.

8.5.2 BSR control element

The BSR MAC control elements are used to report soft buffer size of either a single logical channel group (LCG) or a set of LCGs. An LCG is a group of logical channels identified by a unique 2-bit LCG ID. The mapping of logical channels to an LCG is set up during RRC configuration. There are up to four LCGs that can be set up per UE. The BSR reports aggregate bits across all logical channels of an LCG. As shown in Figure 8.18, there are two formats for the BSR: short/truncated BSRs indicate a buffer size of one LCG, and long BSR indicates buffer sizes of all four LCGs. The buffer size is encoded in 6 bits and includes all data in a logical channel buffer excluding the RLC/MAC headers. The structure of BSR MAC control elements are shown in Figure 8.18.

The UE generates three different types of BSRs as follows:

1. *Regular BSRs* are reported when data becomes available for a logical channel with higher priority than the logical channels for which data exists in the transmission buffer and has already been reported, when a serving cell changes, requiring the UE to report its buffer sizes to the new cell so that uplink resources can be scheduled, or when the timer *RETX_BSR_TIMER* expires and the UE has data available for transmission. This timer is restarted whenever there is an uplink grant and is used for retransmission of a BSR, if there is no allocation for the previous BSR sent by the UE. The regular BSRs can be short or long, depending on what data is available for transmission, and result in generation of a scheduling request at the physical layer.

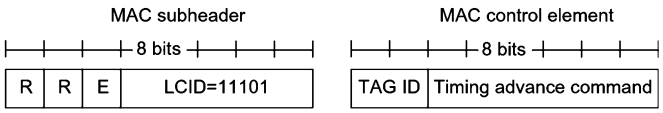


FIGURE 8.17

Uplink timing advance control element [1].

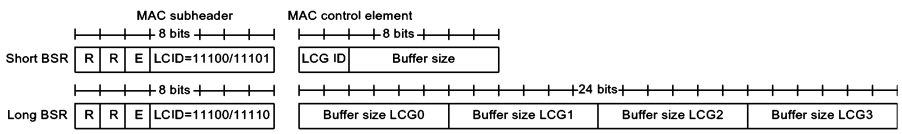


FIGURE 8.18

BSR control elements [1].

2. *Padding BSRs* enable the utilization of bits that would otherwise be used for padding. If the number of padding bits to be sent is equal to or greater than BSR MAC CEs, then BSRs can be sent in lieu of padding. Short, long, or truncated types of BSRs can be sent depending on the size of padding. The truncated BSRs are sent when data is available in more than one LCG, but the size of padding precludes sending long BSRs. In that case, the buffer size of the highest priority LCG is reported.
3. *Periodic BSRs* are sent when the configurable timer *PERIODIC_BSR_TIMER* expires. In this case, short or long BSRs can be sent.

8.5.3 Power headroom control element

The UE power headroom report (PHR) control element is used to report the power headroom available in the UE. The PHR is encoded in 6 bits with a reporting range from −23 dB to +40 dB in steps of 1 dB. Positive values indicate the difference between the maximum UE transmit power and current UE transmit power. Negative values indicate the difference between the maximum UE transmit power and the calculated UE transmit power. The UE transmit power is calculated assuming the UE were to transmit according to the current grant with allocated HARQ and RV configuration. The PHR can be configured either for periodic report or when the downlink path loss changes by a specific amount. For periodic reporting, a report is triggered by the expiration of the *PERIODIC_PHR_TIMER*, which can be configured with values between 10 ms and infinity. For threshold reporting, a PHR is triggered when the path loss changes by *DL_PathlossChange* (1, 3, 6, or infinite dB), provided that the configurable timer *PROHIBIT_PHR_TIMER* has

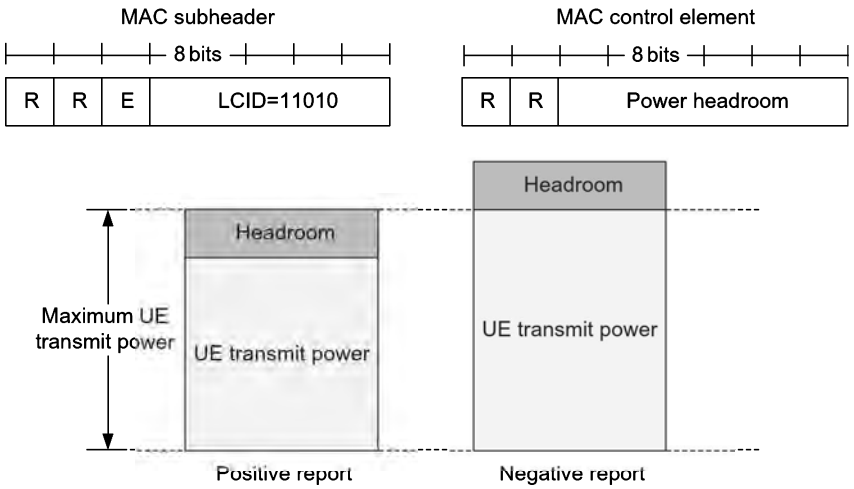


FIGURE 8.19

Power headroom control element [1].

expired. The *PROHIBIT_PHR_TIMER* timer starts when a PHR has been transmitted and can have values between 0 and 1000 ms. The structure of a PHR MAC CE is shown in Figure 8.19, as well as the graphical interpretation of the positive and negative PHRs.

8.5.4 Extended power headroom control element

The extended power headroom MAC control element was specified to support carrier aggregation in LTE Rel-10. In the case of carrier aggregation, the UE reports the power headroom of each component carrier independently. The PHRs provide the serving eNB with information about the difference between the UE maximum power and the estimated transmit power on PUSCH, or the transmit power on PUSCH and PUCCH on the primary carrier. There are two types of PHRs to support simultaneous transmission of PUSCH and PUCCH. Type-1 power headroom is used to report PUSCH transmit power and Type-2 power headroom is used when reporting PUSCH and PUCCH transmit powers. Figure 8.20 shows the order at which the power headroom corresponding to the primary cell and secondary cells are reported in the extended power headroom MAC control element. Note that P_{CMAX_k} in the figure denotes the maximum configured UE transmit power on the kth component carrier. The information fields within the extended power headroom MAC control element are as follows [1]:

- C_i: This bit indicates the presence of a power headroom field for the ith secondary component carrier (secondary cell). The C_i bit is set to one when the power headroom for the ith secondary cell is reported, i.e., the ith active secondary component carrier.
- The reserved bit R is set to zero.
- The information field V indicates whether the reported value of the power headroom is based on a real transmission or a reference. For Type-1 power headroom, $V =$ "0" indicates real transmission on PUSCH and $V = 1$ is indicative of a PUSCH reference format. For Type-2 power headroom, $V =$ "0" indicates real transmission on PUCCH and $V =$ "1" indicates that a PUCCH reference format is used. Furthermore, for both Type-1 and Type-2 power headroom, $V =$ "0" indicates the presence of an octet containing the associated P_{CMAX_k} field, and $V =$ "1" indicates that the octet containing the associated P_{CMAX_k} field is omitted.
- Power Headroom: This 6-bit field indicates the power headroom value.
- P: This field indicates whether the UE applies power backoff due to power management. The UE sets P to one if the corresponding P_{CMAX_k} field might have had a different value if no power backoff due to power management had been applied.
- P_{CMAX_k}: This field indicates the maximum configured UE transmit power on the kth component carrier that is used for calculation of the preceding power headroom value.

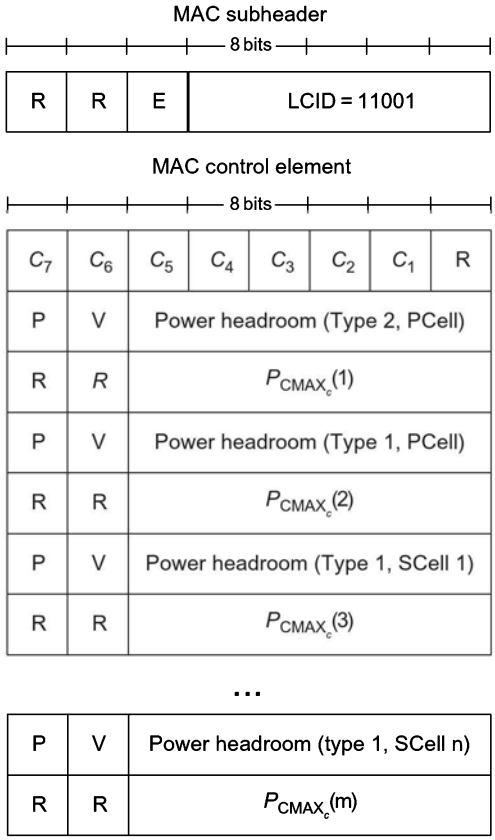


FIGURE 8.20

Extended power headroom control element [1].

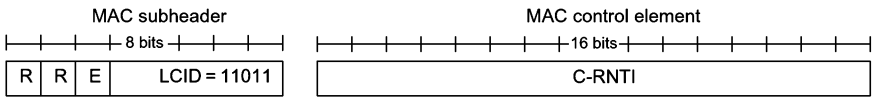


FIGURE 8.21

C-RNTI control element [1].

8.5.5 C-RNTI control element

The C-RNTI MAC control element is used to identify a UE when sending information over CCCH, and for the UE to transmit a dedicated C-RNTI during the random access procedure for the purpose of contention resolution. This CE has a fixed 16-bit size and consists of a single field, as shown in Figure 8.21.

8.5.6 Activation/deactivation control element

LTE Rel-10 specified an activation/deactivation mechanism for secondary cells, i.e., activation/deactivation does not apply to the primary cell, in order to enable reasonable UE battery consumption when carrier aggregation is configured for a UE. When a secondary cell is deactivated, the UE does not need to receive PDCCH or PDSCH on that secondary component carrier and cannot transmit in the corresponding uplink of that cell. However, the UE is required to perform channel quality indicator (CQI) measurements on the deactivated secondary cell. When a secondary cell is activated, the UE is required to monitor PDCCH in order to receive PDSCH and to perform CQI measurements on that cell.

If the UE is configured with one or more secondary cells, the network may activate and deactivate the configured secondary cells. The primary cell is always activated. The network activates and deactivates the secondary cells by sending the activation/deactivation MAC control element. Furthermore, the UE maintains an *sCellDeactivationTimer* per configured secondary cell and deactivates the associated secondary cell upon expiration of its timer. The same initial timer value applies to each instance of the *sCellDeactivationTimer* and it is configured by RRC sublayer. The configured secondary cells are initially deactivated upon addition and after a handover.

The activation/deactivation MAC control element is identified by a MAC PDU subheader with a unique LCID, as shown in Figure 8.22. It has a fixed size and consists of a single octet containing a 7-bit carrier activation/deactivation bitmap and one reserved bit which is set to zero. If the *i*th secondary cell is configured, then the C_i bit in this MAC control element indicates activation of the *i*th secondary cell when set to one.

8.5.7 MCH scheduling information control element

The structure of MCH scheduling information MAC control element is illustrated in Figure 8.23. This CE is transmitted for each MCH to inform the MBMS-capable UEs of scheduled data transmission on MTCH and has a variable size. For each MTCH the information fields include a 5-bit LCID which indicates the LCID of the MTCH and an 11-bit stop MTCH indicating the ordinal number of the subframe within the MCH scheduling period, counting only the subframes allocated to the MCH, where the corresponding MTCH stops, with the value of zero corresponding to the first subframe. The special stop MTCH value 2047 indicates that the corresponding MTCH is not scheduled. The range of values between 2043–2046 is reserved.

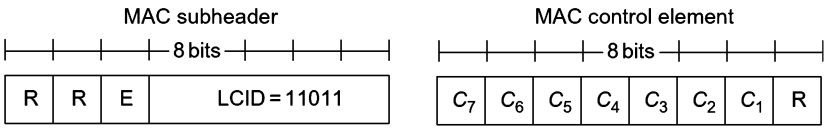


FIGURE 8.22

Activation/deactivation control element [1].

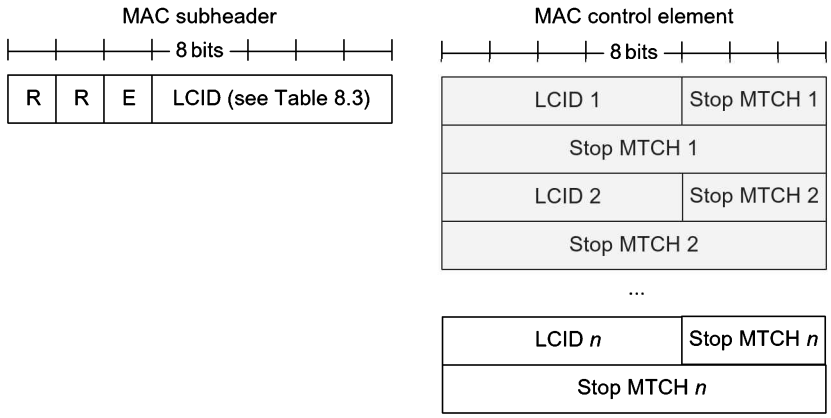


FIGURE 8.23

MCH scheduling information control element [1].

Table 8.3 Values of LCID for MBMS scheduling control element [1]

Index	LCID Values
00000	MCCH (assigned to MTCH, if there is no MCCH on MCH)
00001–11100	MTCH
11101	Reserved
11110	MCH scheduling information
11111	Padding

The MCH may be transmitted in subframes that are configured by an upper layer for MCCH or MTCH transmission. The transmission of an MCH occurs in a set of subframes defined by a *PMCH-Config* parameter. An MCH scheduling information MAC control element is included in the first subframe allocated to the MCH within the MCH scheduling period to indicate the position of each MTCH and unused subframes on the MCH. The UE assumes that the first scheduled MTCH starts immediately after the MCCH or the MCH scheduling information MAC control element, if the MCCH is not present, and the other scheduled MTCH(s) start immediately after the previous MTCH, at the earliest in the subframe where the previous MTCH stops [1] (Table 8.3).

8.5.8 UE contention resolution identity control element

The UE contention resolution identity MAC control element is used during the RACH procedure to resolve possible contention among multiple UEs that are trying to simultaneously access the network. This control element has a fixed 48-bit size and consists of a single field, as shown in Figure 8.24, containing the uplink CCCH SDU.

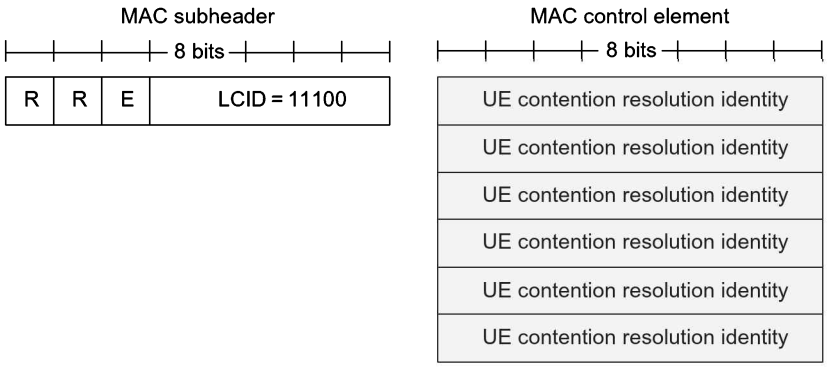


FIGURE 8.24

UE contention resolution identity control element [1].

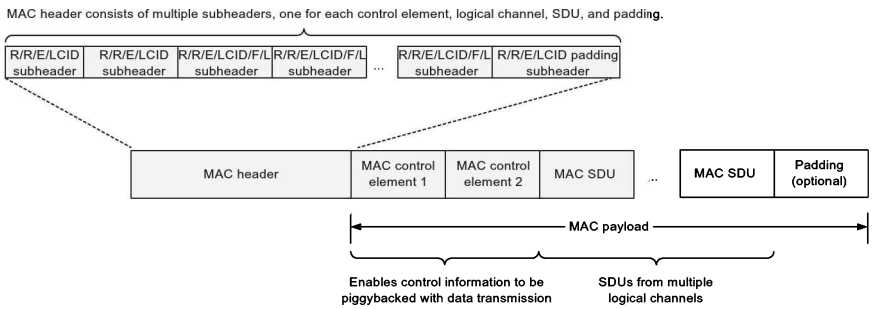


FIGURE 8.25

MAC PDU formation [1].

8.6 MAC PDU packet formats

A MAC PDU consists of a MAC header, zero or more MAC SDUs, zero or more MAC control elements, and optional padding. The MAC header and MAC SDUs may have variable sizes. As shown in Figure 8.25, a MAC PDU header consists of one or more MAC PDU subheaders where each subheader corresponds to a MAC SDU, a MAC control element, or padding. A MAC PDU subheader consists of the six header fields R/R/E/LCID/F/Length except for the last subheader in the MAC PDU and for fixed-size MAC control elements. The last subheader in the MAC PDU and subheaders for fixed-size MAC control elements as well as the subheader for padding consist of the four header fields R/R/E/LCID. Note that a MAC PDU is a bit string that is byte-aligned.

The structure of the MAC PDU used on UL-SCH and DL-SCH is shown in Figure 8.25. Each MAC PDU corresponds to one transport block and consists of a header, control elements, logical channel SDUs (corresponding to RLC PDUs), and

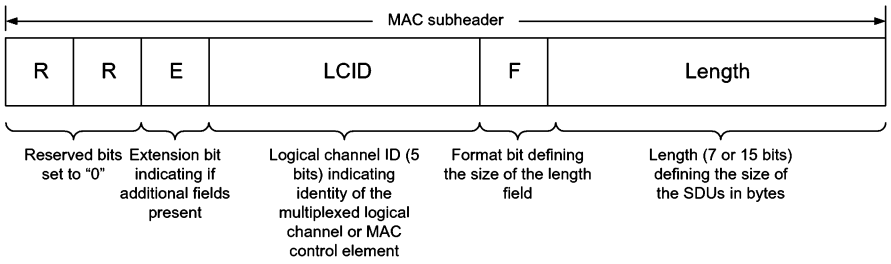


FIGURE 8.26

Structure of a MAC header [1].

optional padding. MAC control elements are used to piggyback control information, such as buffer status or power headroom reports. The MAC header consists of several subheaders. There is one subheader for each MAC control element and each MAC SDU in the PDU. In addition, there can be a subheader to specify padding. For the control element, the subheader identifies the specific control element being sent. For the MAC SDUs, the logical channel number and the size of the SDU are specified. All subheaders are grouped at the start of the MAC PDU in the MAC header. These are followed by the control elements, which are then followed by the MAC SDUs. Any padding occurs at the end of the PDU unless only one or two bytes of padding are required. Then one or two bytes of fixed-size padding subheaders can be included before the subheader corresponding to the first MAC SDU.

The subheader formats for MAC PDUs are shown in Figure 8.26. There are two formats: [R/R/E/LCID] which is used for fixed-length MAC SDUs and MAC control elements; and [R/R/E/LCID/F/Length] that is used for variable length MAC SDUs. The size of the first format is one byte. The size of the second format is either two or three bytes. In each case, the two R bits are reserved bits and are set to "0." The E bit is used to indicate if there are more subheader fields in the MAC header. A value of "1" denotes that another subheader would follow. A value of "0" indicates that a control element or MAC SDU starts at the next byte. The LCID field indicates either the logical channel number associated with the MAC SDU or, in the case of a control element, the specific type of MAC control element. The length field can be either 7 or 15 bits long and specifies the length of the MAC SDU in bytes. The size of the Length field is determined by the F field, which is "0" for 7 bits (128 bytes or less), or "1" for 15 bits.

8.7 MAC scheduling services

The MAC scheduler is responsible for scheduling the transfer of user data and control signaling via transport blocks in downlink and uplink subframes over the LTE air-interface. The MAC scheduler assigns the downlink and uplink physical layer resources used by the eNB and UEs within the cell. The scheduling decisions are based on information from various sources within the eNB radio

protocol stack and the UE. In the downlink, the MAC scheduler allocates resources which are used by the eNB MAC sublayer to send transport blocks to specific UEs via DL-SCH. These downlink transport blocks contain:

- Upper-layer data: All upper-layer data, whether this is user data or signaling messages, are fed into the MAC sublayer through logical channels. Whenever one of these logical channels has any data to send over the LTE air-interface, a buffer status indication is provided to the MAC scheduler.
- MAC sublayer data: This can either take the form of a HARQ retransmission or a MAC control element.

In the above cases, the MAC sublayer will provide an indication to the scheduler that it has data ready to send over the LTE air-interface. In the uplink, the MAC scheduler allocates resources to specific UEs which enable them to construct and send transport blocks to the eNB via UL-SCH. Each UE must request the uplink resources (except when semi-persistent scheduling is used) needed to send a transport block. This can be done in a number of ways. When a UE does not have an active connection to a cell, it performs the random access procedure to attach (or reattach) to the cell. When the UE has uplink control resources on the PUCCH physical channel, it sends a scheduling request indication to request uplink resources. When the UE already has resources to send data in the uplink on the UL-SCH, it may indicate to the scheduler that it has more data to send, and therefore requires more uplink resources. The UE does this by sending a BSR MAC control element. Based on HARQ feedback at the eNB, the scheduler might need to allocate uplink resources for HARQ retransmissions by the UE [12].

The output of the scheduling algorithm is a scheduling assignment per uplink and downlink subframe, which is signaled to UEs on the PDCCH. The downlink assignment information will be used by the eNB MAC sublayer to generate the transport blocks which are passed to the eNB physical layer and subsequently sent over the air. The uplink assignment information indicates which UEs have been allocated uplink resources within the subframe. It is also used by the eNB MAC sublayer to identify uplink transmissions in the corresponding uplink subframe. Before any user data can be transferred over the air, the scheduler needs to be configured with a cell-specific configuration, a UE-specific configuration, and finally a bearer-specific configuration for each active UE bearer. Reconfiguration of each UE and/or bearer is also possible. In order to make full use of the information available within the MAC scheduler, certain event-triggered or periodic feedback information is provided to higher layers. Based on this information, the higher layers are able to reconfigure the operation of the cell to provide a more efficient use of the air-interface. For monitoring purposes, the MAC scheduler may also output detailed information about subframe utilization. This may include but is not limited to the following: subframe map showing utilized physical resource blocks (PRBs); a percentage of utilized versus unutilized PRBs; and the number of logical channels/MAC control elements which have not been scheduled.

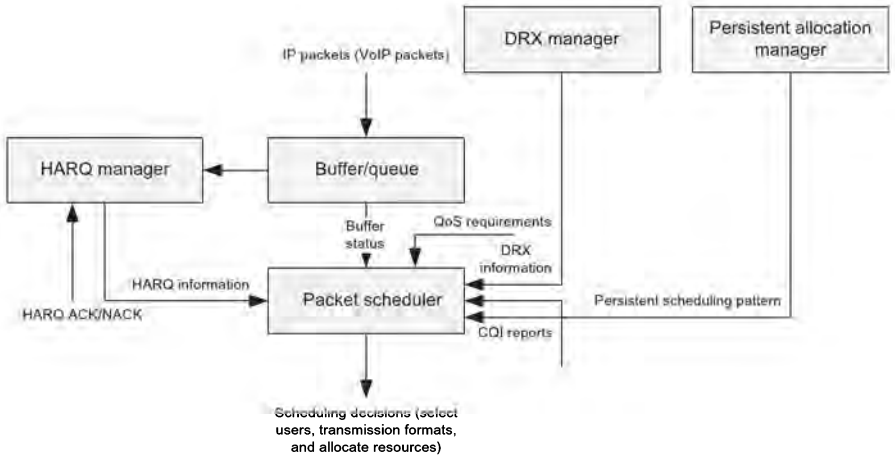


FIGURE 8.27

Main entities involved in packet scheduling in the eNB [15–17].

The scheduler in the eNB distributes the available radio resources in one cell among the UEs, and among the radio bearers of each UE. The details of the scheduling algorithm are left to the eNB implementation, but the signaling to support the scheduling is standardized. In principle, the eNB allocates downlink or uplink radio resources to each UE based on the downlink data buffered in the eNB and the BSRs received from the UE. In this process, the eNB considers the QoS requirements of each configured radio bearer and selects the size of the MAC PDU. The typical mode of scheduling is dynamic scheduling by means of downlink assignment messages for the allocation of downlink transmission resources and uplink grant messages for the allocation of uplink transmission resources; these are valid for specific subframes [12].

Figure 8.27 illustrates the interaction between the main entities involved in eNB scheduling services. These entities include packet scheduler, persistent allocation manager, HARQ manager, and DRX manager. Each newly arrived VoIP packet in the eNB buffer is associated with a discard timer. The discard timer specifies the maximum time a packet is allowed to be buffered in the eNB before it should be dropped. The HARQ manager deals with the new transmissions and pending retransmissions. The DRX manager is responsible for informing the packet scheduler when a specific UE is awake and can receive the L1/L2 control signaling, and thus can be scheduled dynamically. Besides informing each UE of the persistent resource allocation pattern by RRC signaling, the persistent allocation manager also forwards the persistent allocation patterns to the packet scheduler. The packet scheduler is the central controlling entity and makes scheduling decisions based on the buffer status, HARQ information, DRX information, QoS requirements, and persistent resource allocation patterns [13].

These messages also indicate whether the scheduled data is a fresh transmission of a new transport block or a retransmission, using a single-bit NDI. If the

value of the NDI is changed relative to its previous value for the same HARQ process, the transmission is the start of a new transport block. These messages are transmitted on PDCCH using a C-RNTI to identify the addressee (UE). This type of scheduling is efficient for service types, such as transport control protocol (TCP) or the signaling radio bearers (SRBs), in which the traffic is bursty and dynamic. The dynamic uplink transmission resource grants are valid for specific single subframes for initial transmissions, although they may also imply a resource allocation in later subframes for HARQ retransmissions [12].

8.7.1 Dynamic and semi-persistent scheduling

The UE may be dynamically scheduled on a per subframe basis, with the control information signaled on the PDCCH. In this case, the UE will be addressed using the C-RNTI. However, the UE may also receive a semi-persistent grant or allocation where the UE is addressed using the SPS-C-RNTI. In this case, the scheduling control information is signaled once via the PDCCH. The UE can then transmit/receive periodically following a pattern defined by the downlink semi-persistent scheduling interval parameter using the same configuration until modified or released. When a dynamic allocation or grant is received that corresponds to a subframe where an semi-persistent scheduling scheduled transmission would occur, the dynamic grant takes precedence. Additionally, any retransmissions would be scheduled dynamically via the PDCCH.

VoIP is characterized by small packet size and very stringent delay and jitter requirements, and a large number of simultaneous users. In LTE, the scheduling of the data traffic is based on the shared-channel transmission and the channel-dependent resource allocation. The eNB allocates radio resources in units of resource blocks to UEs, and sends the scheduling information to each UE on PDCCH. Such dynamic channel-dependent scheduling can exploit the selectivity in both the time and frequency domain, and significantly improve the system throughput. However, applying dynamic scheduling directly for VoIP packets is not desirable. This is mainly due to the small packet size and the constant inter-arrival time of VoIP packets, which results in large control overhead as control information should be transmitted per TTI. Many UEs may be scheduled simultaneously, so the limited resources available on the PDCCH for control signaling become a bottleneck for the VoIP capacity.

The semi-persistent scheduling is a feature that significantly reduces control channel overhead for applications that require persistent radio resource allocations, such as VoIP. In LTE, both downlink and uplink radio resources are fully scheduled since downlink and uplink traffic channels are shared channels. This means that the PDCCH must provide access grant information to indicate which users should decode PDSCH in each subframe, and which users are allowed to transmit on PUSCH in each subframe. Without semi-persistent scheduling, every downlink or uplink resource block allocation must be granted via an access grant

message on the PDCCH. This is sufficient for most bursty best effort types of applications, which generally have large packet sizes and thus typically only a few users must be scheduled in each subframe. However, for applications that require persistent allocations of small packets, such as VoIP, the access grant control channel overhead can be greatly reduced with semi-persistent scheduling.

Semi-persistent scheduling introduces a persistent resource allocation that a user should expect on the downlink or can transmit on the uplink. There are many different ways in which semi-persistent scheduling can set up persistent allocations. Note that speech codecs typically generate a speech packet every 20 ms. In LTE, HARQ RTT is 8 ms, which means retransmissions of resource blocks that have failed to be decoded can occur every 8 ms. Once speech packets stop arriving (i.e., silence period), these radio resources can be reassigned to other users. When the user begins talking again, a new semi-persistent scheduling set of radio resources would be assigned for the duration of the new talk spurt. Note that dynamic scheduling of best effort data can occur on top of semi-persistent scheduling, but the semi-persistent scheduling allocations would take precedence over any scheduling conflicts.

In wireless networks, persistent scheduling takes advantage of the traffic characteristics of the application (e.g., VoIP) to increase the number of users running that application in the system. The periodic nature of the packet arrivals from a VoIP source allows the needed resource to be allocated persistently for the period of an active talk spurt. Consequently, base stations can send the allocation information once at the beginning of a talk spurt and avoid sending different allocation information for each subsequent packet. Significant amounts of radio resources that would otherwise be occupied by unnecessary overhead can be used to accommodate VoIP packets from other users. Because VoIP packets are usually small in size, the savings from reduced overhead can significantly increase the overall system capacity. The capacity gain from persistent scheduling, however, can be compromised significantly due to dynamic link adaptation, which is a very common technique in wireless systems to adapt to dynamic variation in wireless channel condition. The underlying concept of persistent and dynamic scheduling is illustrated in Figure 8.28.

When semi-persistent scheduling is enabled by the RRC sublayer, the following information is provided: semi-persistent scheduling C-RNTI; uplink semi-persistent

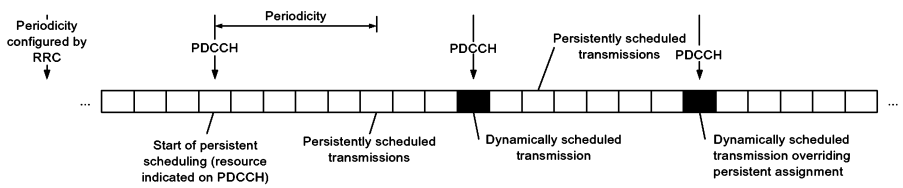


FIGURE 8.28

Illustration of an example of semi-persistent and dynamic scheduling [10].

scheduling interval; the number of empty transmissions before implicit release; downlink semi-persistent scheduling interval; and the number of configured HARQ processes for semi-persistent scheduling. When semi-persistent scheduling for uplink or downlink is disabled by the RRC sublayer, the corresponding configured grant or configured assignment is discarded. Semi-persistent scheduling is supported on the PCell only, and is not supported for relay node communication with the E-UTRAN in combination with a relay node subframe configuration.

After a semi-persistent downlink assignment is configured for a UE, the UE's downlink assignments occur every Nth subframe in a subframe whose number n_{subframe} satisfies the following equation $(10\text{SFN} + n_{\text{subframe}}) = [(10\text{SFN}_{\text{start-time}} + n_{\text{subframe}_{\text{start-time}}}) + NT_{\text{semiPersistSchedIntervalDL}}]\text{mod}10240$, where $\text{SFN}_{\text{start-time}}$ and $n_{\text{subframe}_{\text{start-time}}}$ are the SFN and subframe numbers at which the configured downlink assignment was initiated/reinitiated. Following configuration of a semi-persistent scheduled uplink grant for the UE, the UE will sequentially have uplink grants every Nth subframe for which $(10\text{SFN} + n_{\text{subframe}}) = [(10\text{SFN}_{\text{start-time}} + n_{\text{subframe}_{\text{start-time}}}) + NT_{\text{semiPersistSchedIntervalUL}} + \mathit{Subframe} - \mathit{Offset}(N \bmod 2)]\text{mod}10240$, where $\text{SFN}_{\text{start-time}}$ and $n_{\text{subframe}_{\text{start-time}}}$ are the SFN and subframe number at which the configured uplink grants were initialized/reinitialized. The UE is required to release the configured uplink grant immediately after a certain number of consecutive new MAC PDUs, each containing zero MAC SDUs, have been provided by the multiplexing and assembly entity on the semi-persistent scheduled resource. The retransmissions for semi-persistent scheduling can continue after releasing the configured uplink grant [1].

8.7.2 Proportional fair scheduling algorithm

Proportional fair is a commonly-used scheduling algorithm, which is based on maintaining a balance between two competing interests: maximizing the overall system throughput while at the same time providing all users with a minimum level of service. This is done by assigning each data flow a data rate or a scheduling priority (depending on the implementation) that is inversely proportional to its anticipated resource consumption. The proportional fair (PF) scheduling algorithm calculates a metric for all active users in a given scheduling interval. The user with the highest metric is allocated the resource available in the given interval, while the metrics for all users are updated before the next scheduling interval, and the process repeats. To adapt this simple algorithm to orthogonal frequency division multiple access systems, the definition of scheduling interval and scheduling resource are extended to encompass a two-dimensional OFDMA frame structure. For OFDMA systems, the scheduling interval is typically a radio frame and multiple users may be allocated in the same radio frame. Therefore, in the simplest extension to OFDMA systems, two modifications are made to the PF scheduling algorithm: (1) frames are equally partitioned into regular, fixed scheduling resources that are scheduled sequentially until all available resources are

assigned; (2) the metric is updated after scheduling each partition. Note that the number of resources eventually allocated to a user depends on the metric update process and does not preclude a single user from obtaining multiple or all the resources in a frame. For system simulations with the assumption of a fixed overhead allowing up to $N_{\text{Partitions}}$ resource partitions, each partition assignment should be considered as a separate packet transmission. The number of partitions $N_{\text{Partitions}}$, the time constant of the filter used in the metric computation, and the number of active users are simulation parameters. At any scheduling instant t, the scheduling metric $M_i(t)$ for subscriber i used by the PF scheduler is given by the following expression:

$$M_i(t) = \frac{R_{\text{instantaneous}_i}(t)}{[R_{\text{average}_i}(t)]^{\alpha}} \tag{8.3}$$

where $R_{\text{instantaneous}_i}(t)$ is the data rate that can be supported at scheduling instant t for the ith user. $R_{\text{instantaneous}_i}(t)$ is a function of the CQI feedback, and consequently a function of the MCS that can meet the packet error rate requirement. Parameter $R_{\text{average}_i}(t)$ denotes the throughput function smoothed by a low-pass filter at the scheduling instant t for user i. The parameter α is a fairness exponent factor with a default value of one. For a scheduled subscriber $R_{\text{average}_i}(t)$ is computed as follows:

$$R_{\text{average}_i}(t) = \frac{1}{N_{\text{PF}}} R_{\text{instantaneous}_i}(t) + \left(1 - \frac{1}{N_{\text{PF}}}\right) R_{\text{average}_i}(t-1) \tag{8.4}$$

For an unscheduled subscriber, the throughput function is given by:

$$R_{\text{average}_i}(t) = \left(1 - \frac{1}{N_{\text{PF}}}\right) R_{\text{average}_i}(t-1) \tag{8.5}$$

The latency scale of the PF scheduler N_{PF} is given by:

$$N_{\text{PF}} = \frac{T_{\text{PF}} N_{\text{Partitions}}}{T_{\text{Frame}}} \tag{8.6}$$

where T_{PF} represents the latency time scale in units of seconds and T_{Frame} is the frame duration of the system. In some implementations, the scheduler may give priority to HARQ retransmissions.

8.8 RNTI mapping

The users in LTE radio access networks are assigned temporary identifiers for identification while protecting their privacy and confidentiality. Depending on the state of the UE, different types of temporary identifiers are used. Table 8.4 summarizes the RNTIs and their usage in a LTE radio access network. Each RNTI is a 16-bit

string that is assigned to a UE depending on the state of the UE in the LTE radio access network. For example, the RA-RNTI is used on PDCCH when RAR messages are transmitted. It unambiguously identifies which time-frequency resource was utilized by the UE to transmit the random access preamble. The DCI formats 1A and 1C with CRCs scrambled with RA-RNTI allocate DL-SCH resources to RAR messages. There is a one-to-one mapping between the RA-RNTI and the time-frequency resource used by the UE when transmitting the random access preamble. The one-to-one association allows the eNB to address RAR messages to a specific UE. The Msg3 acronym in Table 8.4 denotes the *RRC Connection Request* message transmitted on UL-SCH containing a C-RNTI MAC control element or CCCH SDU, submitted from the upper layer and associated with the UE contention resolution identity, as part of a random access procedure [10–12].

The P-RNTI is used during the paging procedure, where the UE searches for DCI formats 1A and 1C whose CRC bits are scrambled with P-RNTI. These DCI

Table 8.4 Various RNTIs and Their Usage in LTE [1]

RNTI	Usage	Transport Channel	Logical Channel
Paging RNTI (P-RNTI)	Paging and system information change notification	PCH	PCCH
SI-RNTI	Broadcast of SI	DL-SCH	BCCH
MBMS RNTI (M-RNTI)	MCCH information change notification	N/A	N/A
RA-RNTI	Random Access Response	DL-SCH	N/A
Temporary C-RNTI	Contention resolution (when no valid C-RNTI is available)	DL-SCH	CCCH
Temporary C-RNTI	Msg3 transmission (RRC connection request)	UL-SCH	CCCH, DCCH, DTCH
Cell RNTI (C-RNTI)	Dynamically scheduled unicast transmission	DL-SCH, UL-SCH	DCCH, DTCH
C-RNTI	Triggering of PDCCH ordered random access	N/A	N/A
SPS C-RNTI	Semi-persistently scheduled unicast transmission (activation, reactivation, and retransmission)	DL-SCH, UL-SCH	DCCH, DTCH
SPS C-RNTI	Semi-persistently scheduled unicast transmission (deactivation)	N/A	N/A
TPC-PUCCH-RNTI	Physical layer uplink power control	N/A	N/A
TPC-PUSCH-RNTI	Physical layer uplink power control	N/A	N/A

formats allocate DL-SCH resources to the PCCH and PCH, i.e., the UE proceeds to decode a paging message following detection of a DCI format whose CRC bits are scrambled with P-RNTI. The paging message may include paging information as well as system information change notification, earthquake and tsunami warning system notification, and commercial mobile alert system.

The SI-RNTI is used during the transmission of SIBs on DL-SCH. The UE searches for DCI formats 1A and 1C whose CRC bits are scrambled with SI-RNTI. DCI formats 1A and 1C allocate DL-SCH resources to BCCH. Upon detection of these DCI formats, the UE proceeds to decode the system information message.

The M-RNTI is used during transmission of an MCCH change notification. The UE searches for a DCI format 1C whose CRC is scrambled with M-RNTI. This DCI format includes a bit string of 8 bits, where each bit corresponds to one of the MBSFN areas listed in SIB13. A change in MCCH is signaled by setting the corresponding bit to one.

The temporary C-RNTI is allocated to the UE within the RAR message. The temporary C-RNTI is used for contention resolution when a C-RNTI has not yet been assigned and specifically in the transmission of the RRC connection request on CCCH during initial access. After transmitting Msg3 in the uplink, the UE searches for DCI formats 1A or 1C whose CRC bits are scrambled with the UE's temporary C-RNTI. The successful detection of the temporary C-RNTI is considered success of the contention resolution and assignment of a C-RNTI to the UE. The temporary C-RNTI is also used for scrambling of PUSCH during initial Msg3 transmission on CCCH.

The C-RNTI is assigned to the UE following successful completion of the random access procedure. The C-RNTI is reassigned during the handover procedure by specifying a new value within the *RRC Connection Reconfiguration* message. The C-RNTI is used to identify the UE in all downlink and uplink control and data transmissions when the UE is in the RRC_CONNECTED state. The UE uses the same C-RNTI in all serving cells when supporting carrier aggregation.

The SPS-C-RNTI is allocated to a UE using an *RRC Connection Setup* or an *RRC Connection Reconfiguration* message, when the resources are going to be allocated for more than a single subframe. The transmit power control (TPC) PUSCH/PUCCH RNTI is used for power control of PUSCH/PUCCH. The CRC bits of DCI formats 3 and 3A are scrambled with TPC-PUSCH-RNTI/TPC-PUCCH-RNTI to convey the power control commands to the UE. TPC-PUSCH-RNTI/TPC-PUCCH-RNTI can be allocated using either the *RRC Connection Setup* or *RRC Connection Reconfiguration* messages [10]. The mapping of different types of RNTI to logical channels is depicted in Figure 8.29.

8.9 Random access procedure

The UE must be synchronized in the uplink before it can send signaling or data. For the uplink, the UEs within a cell's coverage area are physically distributed at different distances from the eNB location. As a result, each UE will experience a

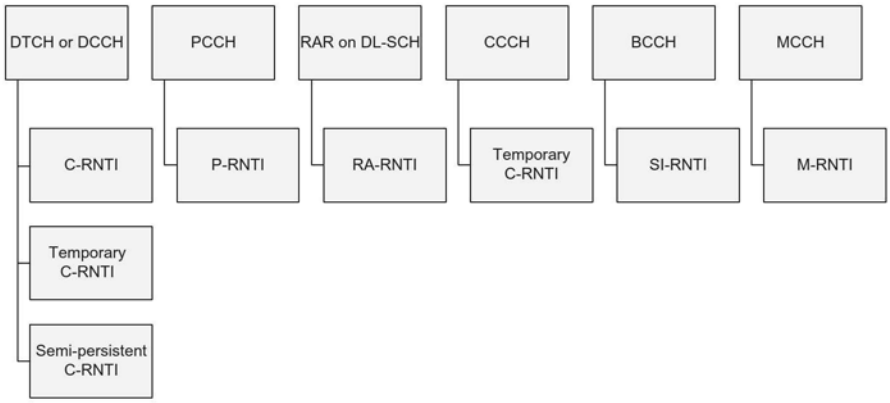


FIGURE 8.29

RNTI mapping to control or traffic logical channels.

different propagation delay when accessing the network. Without synchronization, subframes transmitted from different UEs would likely overlap and interfere with each other. To overcome this problem, the UE performs uplink synchronization procedures on RACH during initial access to the network. In LTE, random access is used for several purposes including initial access when establishing a radio link, i.e., transitioning from RRC_IDLE to RRC_CONNECTED state, reestablishing the radio link after a radio link failure, performing handover when uplink synchronization, needs to be established with the target cell, establishing uplink synchronization if uplink or downlink data arrives when the terminal is in the RRC_CONNECTED state and the uplink is not synchronized, positioning during RRC_CONNECTED state which requires a random access procedure, e.g., when timing advance is needed for UE positioning, and scheduling request if no dedicated scheduling-request resources have been configured on PUCCH.

The process is initiated by sending RACH preambles using physical random access channel (PRACH). The procedure may be contention-based or contention-free. In the contention-based procedure, the UE selects a preamble from a set containing two groups of preamble sequences. If two UEs select the same preamble in the same RACH resource, a collision can result that must be resolved before either UE can access the system. The contention-based RACH is generally used during initial access or if the UE has lost uplink synchronization. In contention-free synchronization, the UE uses a dedicated preamble and it is likely to be already connected to the network and must resynchronize following a handover to a new cell.

Figure 8.30 shows the RACH process at a high level. During the first step the UE transmits a preamble randomly selected from 64 possible preamble sequences divided into two groups. The group from which the UE selects the preamble depends on the size of the message that it would like to transmit and on the

Samsung Ex. 1005, Page 421 of 1147

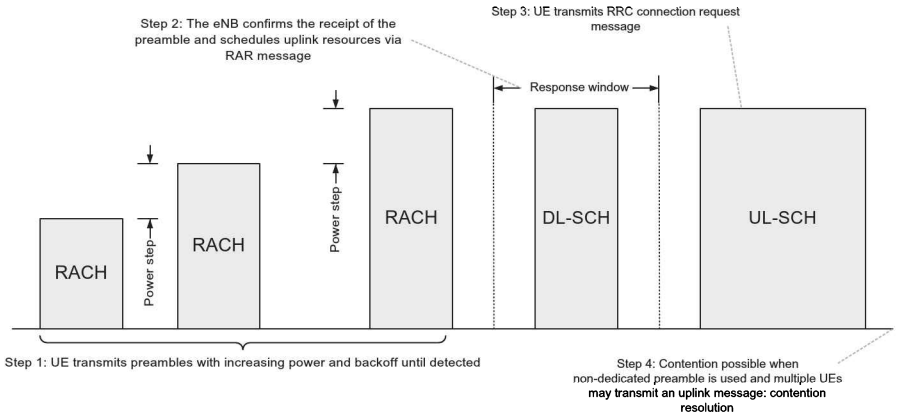


FIGURE 8.30

Illustration of a contention-based random access procedure [21].

calculated path loss to the eNB. Once the preamble is transmitted, the UE monitors the downlink beginning three subframes after the end of the preamble for a duration controlled by a configurable window. If no downlink response is detected, then the preamble is retransmitted with a power that is increased by a configurable parameter. This process continues until the defined maximum number of preamble transmissions is reached. If the preamble is successfully detected by the eNB, then the downlink RAR message is sent by the eNB allocating resources for the UE to transmit its message or enforcing a backoff. In the final step, contention resolution must be performed for a contention-based access attempt. Therefore, as soon as the random access preamble is transmitted and regardless of the possible occurrence of a measurement gap, the UE is required to monitor PDCCH on the primary cell for RAR(s) identified by RA-RNTI during the RAR window (in time) which starts at the subframe that contains the end of the preamble transmission plus three subframes and has the length *ra-ResponseWindowSize* subframes (Figure 8.30) [1].

The UE sets its initial power level to the following value
PREAMBLE_INITIAL_RECEIVED_TARGET_POWER + *DELTA_PREAMBLE*
where *DELTA_PREAMBLE* depends on the type of preamble format. The UE selects the first available RACH resource for the transmission. If there is no response within three subframes + *ra_Response_Window* (defined in number of subframes), the UE assumes the previous transmission was unsuccessful. It selects another RACH preamble, performs the backoff procedure, and transmits again. The power level is increased by *POWER_RAMP_STEP*. This process can be repeated until the *PREAMBLE_TRANSMISSION_COUNTER* limit is reached. In that case, an access failure is declared to the upper layers.

In LTE Rel-10 and when carrier aggregation is supported, the random access procedure is initiated by a PDCCH order or by the MAC sublayer. The random access procedure on a secondary cell is only initiated by a PDCCH order

(i.e., non-contention-based random access procedure). If the UE receives a PDCCH order masked with the UE's C-RNTI, and for a specific serving cell, the UE initiates a random access procedure on that serving cell. For random access on the primary cell, a PDCCH order or RRC message optionally indicates the *ra-PreambleIndex* and the *ra-PRACH-MaskIndex* parameters; whereas for random access on a secondary cell, a PDCCH order indicates the *ra-PreambleIndex* (with a value different from 000000) and the *ra-PRACH-MaskIndex* parameters. For the primary TAG, preamble transmission on PRACH and reception of a PDCCH order are only supported for the primary cell [1]. In LTE Rel-11, the random access procedure is also performed on a secondary cell to establish timing alignment for the corresponding secondary TAG. Furthermore, a non-contention-based random access procedure can be performed for obtaining timing advance alignment for a secondary TAG [8].

8.9.1 Contention-based RACH procedure

The UE selects a preamble from the set of preambles. Two groups of sequences are available for a contention-based RACH procedure defined as Group A and Group B. The size of the Group A preamble set is defined by the *sizeofRA-PreamblesGroupA* parameter. The remaining preambles, up to *numberofRA-Preambles*-1, belong to Group B. Group B preambles are utilized if the required message size exceeds *MESSAGE_SIZE_GROUP_A* and the path loss is less than $P_{\max} - PREAMBLE\_INITIAL\_RECEIVED\_TARGET\_POWER - DELTA\_PREAMBLE\_MSG3 - messagePowerOffsetGroupB$ where $P_{\max}$ is the maximum UE transmit power. This enables the eNB to estimate the size of the message that the UE wants to transmit. If the eNB detects a preamble, it sends a RAR message in the DL-SCH. The MAC PDU containing the RAR message is addressed using the RA-RNTI which is calculated as $RA - RNTI = 1 + t\_id + 10f\_id$ where $0 \le t\_id < 10$ is the index of the first subframe of RACH resource within a frame, and $0 \le f\_id < 6$ is the specified RACH resource within that subframe. This forms a one-to-one mapping between the selected RACH resource and the RA-RNTI. The RAR message contains an identifier to enable identification of the preamble sent by the UE, the absolute timing advance, a temporary C-RNTI, and an uplink grant. If the preamble identifier matches the preamble sent by the UE, the UE applies the timing advance and starts the contention resolution procedure (see Figure 8.31).

During contention resolution, the UE transmits an uplink message on CCCH that includes a terminal identifier. If this matches the identifier used in the response from the eNB, then the UE has successfully completed the contention resolution; otherwise, the UE must restart the random access procedure. Upon successful contention resolution, the temporary C-RNTI assigned as part of the RAR message will become the C-RNTI.

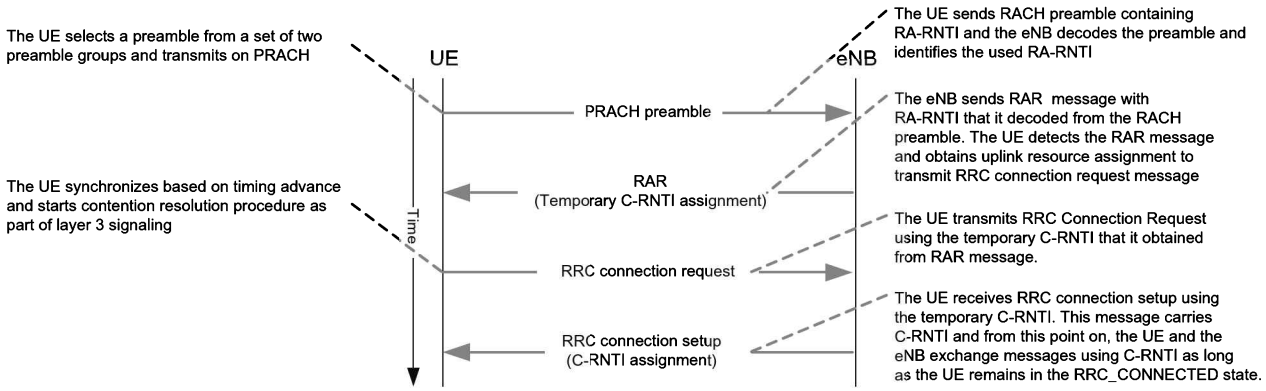


FIGURE 8.31

Breakdown of the contention-based random access procedure [21].

8.9.2 Contention-free RACH procedure

During handovers, the UE must synchronize to the target cell using RACH procedures. However, contention-based random access may result in an unacceptable delay. Therefore, a contention-free procedure is defined. In contention-free RACH, the RRC sublayer defines which RACH preamble sequence and which PRACH resources to utilize. The preamble is dedicated to the UE and is different from those used for Groups A and B in the contention-based process, thus avoiding any possible collision. The PRACH resource to be used for the dedicated transmission is indicated as PRACH Mask, which specifies the subframe number in which the UE can transmit the preamble.

Once the UE transmits the dedicated preamble in the first available RACH resource after receiving the command, it waits for a response in the form of an RAR message from the eNB. The same timing and window specification is enforced in this procedure as those associated with the contention-based procedure. Once the RAR message is received on PDCCH, the UE checks whether the random access preamble identifier (RAPID) contained in the message matches the preamble that was sent and, if so, it applies the timing advance and synchronizes with the eNB.

As mentioned earlier, the MAC PDU that carries the RAR message on PDCCH contains a number of RARs, since multiple UEs may transmit on the uplink RACH resource at the same time. As shown in Figure 8.32, the MAC PDU starts with the MAC header, which contains a number of subheaders. For each RAR message, there is a subheader that identifies the RACH Preamble Identifier that was used for the transmission in uplink. The corresponding RAR payload contains the detailed information on the timing advance, uplink grant, and temporary C-RNTI. The MAC header may also include a backoff indicator (BI) subheader which indexes a backoff ranging from 0 to 960 ms. It is applicable to all the UEs that use a specific RACH resource.

The MAC subheader for the RAR message is 8 bits and can have one of the two formats shown in Figure 8.33. Both RAR formats begin with an E field, which is a single bit that indicates if another subheader is present after the current subheader, and a T field, which is a single bit that defines if this is a subheader that assigns uplink resources or informs all UEs that utilize the RACH resource to

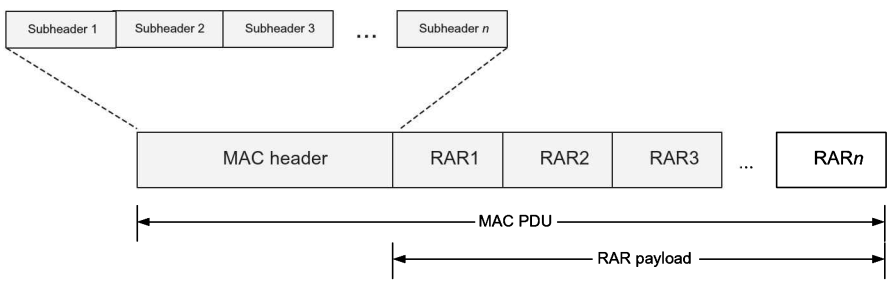


FIGURE 8.32

RAR message structure [1].

backoff. If resources are assigned in a RAR payload, then T is set to "1" and the following field is the 5-bit RAPID value. In this case, there will be an RAR payload to specify the specific grant. If the eNB is enforcing a backoff for all UEs utilizing the RACH resource, then T is set to "0" and the following field consists of two single-bit R (reserved) fields followed by a 4-bit BI that indexes a table of 13 possible backoff periods ranging from 0 to 960 ms.

As shown in Figure 8.34, the RAR payload consists of 48 bits described as follows:

- R: A single reserved bit set to "0."
- TA: The 11-bit timing advance specifies the absolute timing advance that should be applied to the uplink transmissions in units of $16T_s$ (sampling time).
- Uplink grant: 20 bits that specify attributes of the uplink grant. This includes hopping flag (1 bit), defines if frequency hopping is used in the uplink, Fixed-size resource block assignment (10 bits), defines the uplink resource blocks, Truncated MCS (4 bits), TPC command for scheduled PUSCH (3 bits), uplink delay (1 bit), controls which uplink subframe is used for the initial transmission, and CQI request (1 bit), specifies if a CQI report is sent in PUSCH transmission for non-contention-based RACH.
- Temporary C-RNTI: Temporary identifier utilized by the UE until contention resolution is complete.

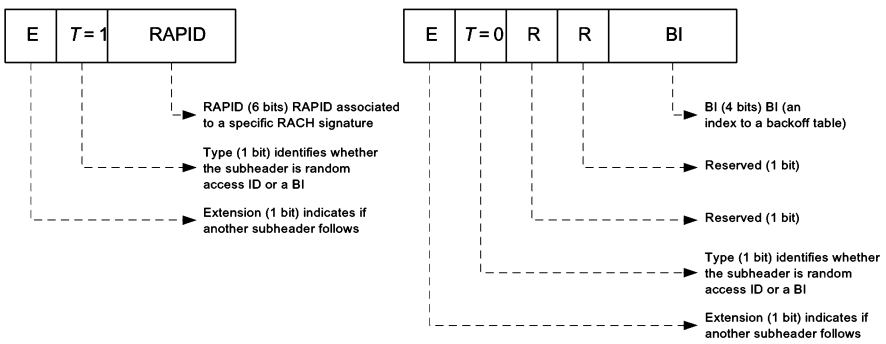


FIGURE 8.33

Structure of RAR subheaders [1].

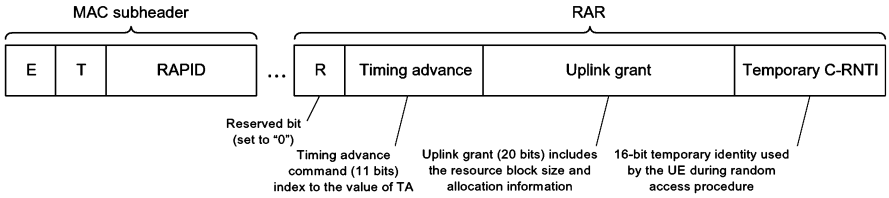


FIGURE 8.34

Structure of RAR payload [1].

References

3GPP Specifications

[1] 3GPP TS 36.321, Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol Specification (Release 11), December 2012.

[2] 3GPP TS 36.322, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) Protocol Specification (Release 11), December 2012.

[3] 3GPP TS 36.323, Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) Specification (Release 11), December 2012.

[4] 3GPP TS 36.331, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC) Protocol Specification (Release 11), December 2012.

[5] 3GPP TS 36.211, Evolved Universal Terrestrial Radio Access (E-UTRA), Physical Channels and Modulation (Release 11), December 2012.

[6] 3GPP TS 36.212, Evolved Universal Terrestrial Radio Access (E-UTRA) Multiplexing and Channel Coding (Release 11), December 2012.

[7] 3GPP TS 36.213, Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures (Release 11), December 2012.

[8] 3GPP TS 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Overall Description, Stage 2 (Release 11), December 2012.

[9] 3GPP TS 36.133, Evolved Universal Terrestrial Radio Access (E-UTRA); Requirements for Support of Radio Resource Management (Release 11), December 2012.

Books

[10] C. Johnson, Long Term Evolution in 10, second ed., Create Space Independent Publishing Platform, July 2012.

[11] S. Sesia, I. Toufik, M. Baker, LTE, the UMTS Long Term Evolution: From Theory to Practice, second ed., John Wiley & Sons, August 2011.

[12] E. Dahlman, S. Parkvall, J. Skold, 4G: LTE/LTE-Advanced for Mobile Broadband, Academic Press, May 2011.

Articles

[13] Y. Fan, P. Lundén, M. Kuusela, M. Valkama, Efficient semi-persistent scheduling for VoIP on E-UTRA downlink, in: Proceedings of the IEEE VTC-2008, Calgary, Canada, September 2008.

[14] Y. Fan, M. Kuusela, P. Lundén, M. Valkama, Downlink VoIP support for Evolved UTRA, in: Proceedings of the IEEE WCNC-2008, Las Vegas, NV, April 2008.

[15] A. Pokhariyal, T.E. Kolding, Performance of downlink frequency domain packet scheduling for the UTRAN long term evolution, in: Proceedings of the IEEE PIMRC-2006, Helsinki, Finland, September 2006.

[16] K.I. Pedersen, et al., Frequency domain scheduling for OFDMA with limited and noisy channel feedback, in: Proceedings of the IEEE VTC-2007, Baltimore, MD, September 2007.

[17] Y. Sun, et al., Multi-user scheduling for OFDMA downlink with limited feedback for Evolved UTRA, in: Proceedings of the IEEE VTC-2006, Montreal, Canada, September 2006.

[18] C.S. Bontu, E. Illidge, DRX mechanism for power saving in LTE, IEEE Commun. Mag. (June 2009).
[19] C. Zhong, et al., A new DRX scheme for LTE-advanced carrier aggregation systems with multiple services, IEEE Vehicular Technology Conference, 2011.
[20] Root of unity, from Wikipedia. <http://en.wikipedia.org/wiki/Root_of_unity>.
[21] LTE system description. <http://www.sharetechnote.com>.

Samsung Ex. 1005, Page 428 of 1147

This page intentionally left blank

CHAPTER

Downlink Physical Layer Functions

9

CHAPTER CONTENTS

9.1 Overview of downlink physical layer processing 402
9.2 Characteristics of wireless channels 404
9.3 LTE/LTE-Advanced operating frequencies and band classes 411
9.4 Principles of OFDM 419
9.4.1 Peak to average power ratio (PAPR) of OFDM signals 426
9.4.2 Effect of non-linearity on OFDM signals 428
9.4.3 Effect of carrier frequency offset on OFDM signals 430
9.4.4 Effect of oscillator phase noise on OFDM signals 434
9.5 Multiple access schemes 437
9.6 Duplex modes 440
9.7 Frame structure 445
9.7.1 Frame structure type 1 446
9.7.2 Frame structure type 2 447
9.7.3 MBSFN subframe structure 452
9.8 Resource structure, allocation, mapping 453
9.8.1 Physical and virtual resource blocks 454
9.8.2 Resource allocation 460
9.8.2.1 Resource allocation type 0 462
9.8.2.2 Resource allocation type 1 463
9.8.2.3 Resource allocation type 2 465
9.8.2.4 PRB bundling 467
9.8.2.5 Resource element group 467
9.9 Time and frequency synchronization 471
9.9.1 Primary synchronization signals 473
9.9.2 Secondary synchronization signals 474
9.10 Downlink reference signals and channel estimation 478
9.10.1 Reference signal design criteria 480
9.10.2 Pilot-Based channel estimation 486
9.10.3 Cell-Specific reference signals 491
9.10.3.1 Received signal strength indicator measurement 496
9.10.3.2 Reference signal received power measurement 496
9.10.3.3 Reference signal received quality measurement 496
9.10.4 UE-Specific reference signals associated with PDSCH 497

LTE-Advanced.
© 2014 Elsevier Inc. All rights reserved.

Samsung Ex. 1005, Page 430 of 1147

9.10.5 CSI reference signals 502
9.10.6 Enhanced PDCCH demodulation reference signals 512
9.10.7 Positioning reference signals 514
9.10.8 MBSFN reference signals 516
9.10.9 Reference signal overhead calculation 518
9.11 Channel coding and modulation 519
9.11.1 Principles of convolutional coding 521
9.11.2 Principles of turbo coding 527
9.11.3 Baseband digital modulation schemes 533
9.11.4 Channel coding, multiplexing, and interleaving 537
9.12 Downlink physical channel processing 546
9.12.1 Physical control format indicator channel 553
9.12.2 Physical broadcast channel 556
9.12.3 Physical hybrid ARQ indicator channel 560
9.12.4 Physical downlink control channel 568
9.12.4.1 PDCCH physical layer processing 568
9.12.4.2 DCI formats 574
9.12.4.3 Common and UE-specific search spaces and blind decoding 576
9.12.5 Enhanced PDCCH 597
9.12.6 Physical downlink shared channel 604
9.13 Principles of multi-antenna transmission 608
9.13.1 Capacity of MIMO channels 610
9.13.2 Instantaneous capacity, ergodic capacity, and outage capacity of MIMO systems 617
9.13.3 Spatial multiplexing and diversity 620
9.13.4 MIMO receivers 624
9.13.4.1 Maximum likelihood receiver 624
9.13.4.2 Successive interference cancelation receiver 624
9.13.4.3 Zero-forcing receiver 625
9.13.4.4 Minimum mean square error receiver 626
9.13.4.5 Maximal ratio combining receiver 627
9.13.4.6 MMSE-interference rejection combining receiver 628
9.13.4.7 Sphere decoding 630
9.13.4.8 CSI feedback 632
9.13.5 Precoding and beamforming 636
9.13.6 Single-user and Multi-user MIMO 642
9.13.7 Collaborative MIMO and collaborative spatial multiplexing 652
9.13.8 Non-linear precoding and dirty-paper coding 654
9.14 LTE/LTE-Advanced downlink multi-antenna transmission schemes 661
9.14.1 Transmit diversity 666
9.14.2 Spatial multiplexing 668
9.14.3 Cyclic delay diversity (CDD) 672
9.14.4 Beamforming and MU-MIMO 674

Samsung Ex. 1005, Page 431 of 1147

9.15 PDSCH transmission modes 679
9.16 OFDM signal generation 686
9.17 Timing advance 686
9.18 Power control and link adaptation 690
9.18.1 Basic concepts 691
9.18.2 Downlink power allocation 698
9.18.3 PUSCH power control 700
9.18.4 PUCCH power control 703
9.18.5 SRS power control 704
9.18.6 PRACH power control 705
9.18.7 Power headroom 706
9.18.8 Link adaptation 707
References 712

This chapter describes the physical layer protocols and functional processing in LTE/LTE-Advanced downlink direction. As shown in Figure 9.1, the physical layer is the lowest protocol layer in baseband signal processing that interfaces with the physical media through which the signal is transmitted and received. The physical layer receives MAC PDUs (transport blocks) and processes them through channel coding, interleaving, baseband modulation, layer mapping for multi-antenna operation, eigen-precoding, resource element, and antenna mapping functions. The choice of appropriate modulation and coding scheme, and multi-antenna transmission mode, are critical to achieve the desired reliability/robustness and system/webuser throughput in mobile wireless data communications. Typical mobile radio channels tend to be dispersive and time-variant and exhibit severe Doppler effects, multipath delay variation, intra-cell and inter-cell interference, and fading.

A good and robust design of the physical layer ensures that the system can normally operate and overcome the above deleterious effects, and can provide the maximum throughput and lowest latency under various operating conditions. The chapters on physical layer in this book are dedicated to systematic design of physical layer protocols and functional blocks of fourth generation cellular systems, the theoretical background on physical layer procedures, and performance evaluation of physical layer components. The theoretical background is provided to make the book self-contained, and to ensure that the reader understands the underlying theory governing the operation of various functional blocks and procedures. Additional references are provided for further study. While the focus is mainly on the techniques that were incorporated in the design of LTE/LTE-Advanced physical layer, the author has attempted to take a more generic and systematic a IMT-Advanced and beyond cellular systems, so that the reader can understand and apply the learning to the design and implementation of any OFDM-based physical layer irrespective of the radio access technology. Physical layer processes both control- and user-plane signals; however,

Samsung Ex. 1005, Page 432 of 1147

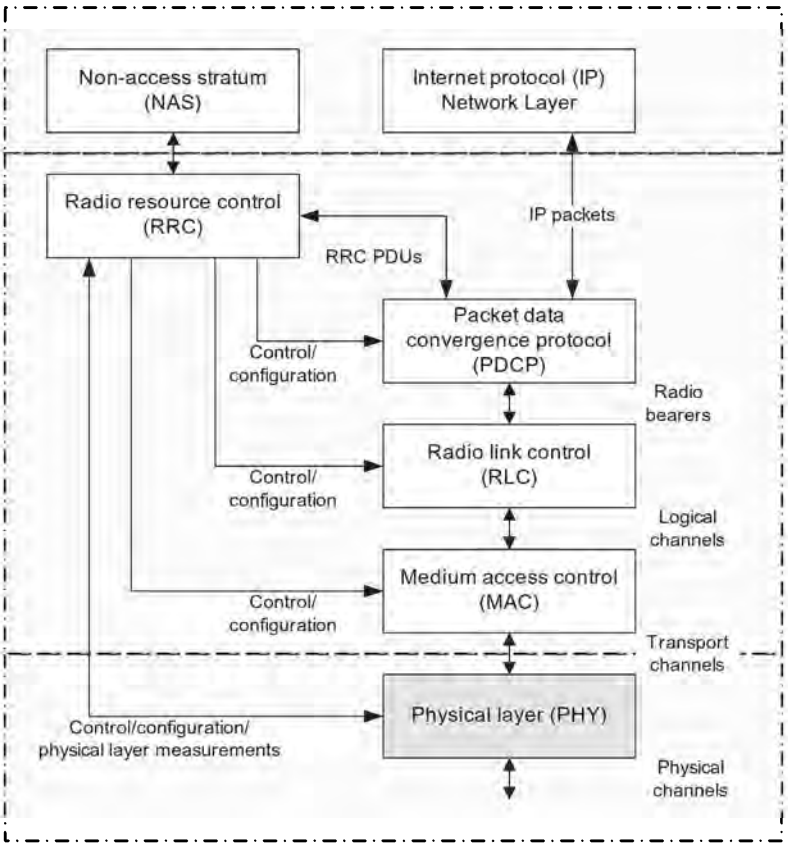


FIGURE 9.1

Physical Layer in LTE Protocol Stack

due to different design requirements and reliability and performance criteria, the procedures tend to be different. We will underline these differences and will describe the methods for selecting the best transmission format for each physical channel.

9.1 Overview of downlink physical layer processing

This section provides an overview of physical layer processing in LTE/LTE-Advanced downlink. The processing steps and the associated control signaling are depicted in Figure 9.2. The main criteria in the design of the LTE/LTE-Advanced physical layer were to increase the application throughput and capacity, reduce access latency, support higher user mobility, minimize intra-cell and inter-cell interference, improve reliability of control and data channel coverage especially

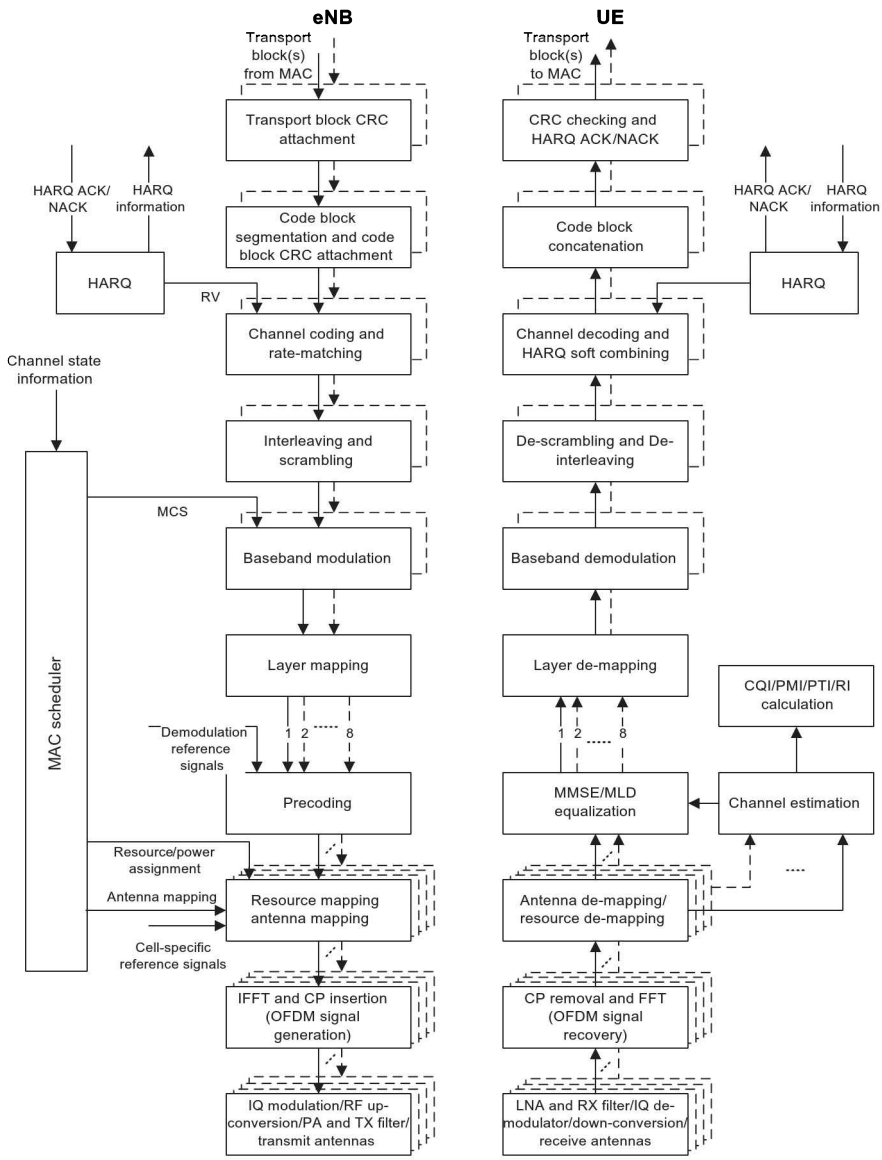


FIGURE 9.2

Downlink physical layer processing in LTE-Advanced [3,156].

at the cell-edge, and reduce the complexity and signaling overhead. The overall performance evaluations provided in Chapter 11 suggest that the desired goals have been successfully achieved. In the next sections of this chapter, we describe in detail the functional blocks and protocols of the downlink physical layer and their interactions based on the order that the information is processed as shown in Figure 9.2. It must be noted that although the overall procedure shown in Figure 9.2 is similar for LTE Rel-8/9 and LTE-Advanced Rel-10/11, the number of transmission layers, transmission modes, and multi-antenna processing are different. These differences will be identified in description of each function in the upcoming sections.

9.2 Characteristics of wireless channels

In a wireless communication system, a signal can travel from the transmitter to the receiver over multiple paths. This phenomenon is referred to as multipath propagation where signal attenuation varies on different paths. This effect, also known as multipath fading, can cause stochastic fluctuations in the received signal's magnitude, phase, and angle of arrival. The propagation over different paths is caused by scattering, reflection, diffraction, and refraction of the radio waves by stationary and moving objects, as well as the medium. It is obvious that different propagation mechanisms result in different channel and path loss models. As a result of wave propagation over multipath fading channels, the radio signal is attenuated due to mean path loss, as well as macroscopic and microscopic fading.

In an ideal free-space propagation model, the attenuation of RF signal energy between the transmitter and receiver follows the inverse-square law. The received power expressed in terms of transmitted power is attenuated proportional to the inverse of $L_s(d)$, which is called free space loss. When the receiving antenna is isotropic, the received signal power can be expressed as follows [19,48,49]:

$$P_{RX} = P_{TX} G_{TX} G_{RX} \left(\frac{\lambda}{4\pi d}\right)^2 = \frac{P_{TX} G_{TX} G_{RX}}{L_s(d)} \tag{9.1}$$

where P_{TX} and P_{RX} denote the transmitted and the received signal power, G_{TX} and G_{RX} denote the transmitting and receiving antenna gains, d is the distance between the transmitter and the receiver, and λ is the wavelength of the RF signal.

Macroscopic fading is caused by shadowing effects of buildings and natural obstructions, and is modeled by the local mean of a fast fading signal. The mean path loss $\overline{L}_p(d)$ as a function of distance d between the transmitter and receiver is proportional to the nth power of d relative to the reference distance d_0. In logarithmic scale, it can be expressed as follows:

$$\overline{L}_p(d) = L_s(d_0) + 10n \log(d/d_0) \quad \text{(dB)} \tag{9.2}$$

The reference distance d_0 corresponds to a point located in the far-field of the antenna typically 1 km for large cells, 100 m for micro-cells, and 1 m for indoor channels. In Eq. (9.2), $\overline{L}_p(d)$ is the mean path loss, which is typically $10n$ dB per decade attenuation for $d \gg d_0$. The value of n depends on the frequency, antenna heights, and propagation environment, and is equal to 2 in free space. Studies show that the path loss $L_p(d)$ is a random variable with log-normal distribution about the mean path loss $\overline{L}_p(d)$. Let $X \sim \mathbb{N}(0, \sigma^2)$ denote a zero-mean Gaussian random variable with standard deviation σ when measured in decibels, then:

$$L_p(d) = L_s(d_0) + 10n \log(d/d_0) + X \quad \text{(dB)} \tag{9.3}$$

The value of X is often derived empirically based on measurements. A typical value for σ is 8 dB. The parameters that statistically describe path loss due to large-scale fading (macroscopic fading) for an arbitrary location with a specific transmitting–receiving antenna separation include the reference distance d_0, the path loss exponent n, and the standard deviation σ of X [19,48,49].

Microscopic fading refers to the rapid fluctuations of the received signal in time and frequency, and is caused by scattering objects between the transmitting and receiving antennas. When the received RF signal is a superposition of independent scattered components plus a line-of-sight (LoS) component, the envelope of the received signal $r(t)$ has a Rician probability distribution function (PDF), and is referred to as Rician fading. As the magnitude of the LoS component approaches zero, the Rician PDF approaches a Rayleigh PDF. Thus:

$$f(r) = \frac{r(K+1)}{\sigma^2} \exp\left[-K - \frac{(K+1)r^2}{2\sigma^2}\right] I_0\left(\frac{2r}{\sigma}\sqrt{K(K+1)/2}\right) \quad r \geq 0 \tag{9.4}$$

where K and $I_0(r)$ denote the Rician factor and zero-order modified Bessel function of the first kind. In the absence of LoS path ($K = 0$), the Rician PDF reduces to Rayleigh distribution.

Time-varying fading due to scattering objects or transmitter/receiver motion results in Doppler spread. The time spreading effect of small-scale or microscopic fading is manifested in the time-domain as multipath delay spread, and in the frequency-domain as channel coherence bandwidth. Similarly, the time variation of the channel is characterized in the time-domain as channel coherence time, and in the frequency-domain as Doppler spread. In a fading channel, the relationship between maximum excess delay time τ_m and symbol time τ_s can be viewed in terms of two different degradation effects, i.e., frequency-selective fading and frequency non-selective or flat fading. A channel is said to exhibit frequency-selective fading if $\tau_m > \tau_s$. This condition occurs whenever the received multipath components of a symbol extend beyond the symbol's time duration. Such multipath dispersion of the signal results in inter-symbol interference (ISI) distortion. In the case of frequency-selective fading, mitigating the distortion is possible because many of the multipath components are separable by the receiver. A channel is said to

exhibit frequency non-selective or flat fading if $\tau_m < \tau_s$. In this case, all received multipath components of a symbol arrive within the symbol time duration; therefore, the components are not resolvable. In this case, there is no channel-induced ISI distortion, since the signal time spreading does not result in significant overlap among adjacent received symbols. There is still performance degradation, because the irresolvable phasor components can add up destructively to yield a substantial reduction in signal-to-noise ratio (SNR). Also, signals that are classified as exhibiting flat fading can sometimes experience frequency-selective distortion.

Figure 9.3 illustrates a multipath-intensity profile $\Lambda(\tau)$ versus delay τ where the term delay refers to the excess delay. It represents the signal's propagation delay that exceeds the delay of the first signal arriving at the receiver. For a typical wireless communication channel, the received signal usually consists of several discrete multipath components. The received signals are composed of a

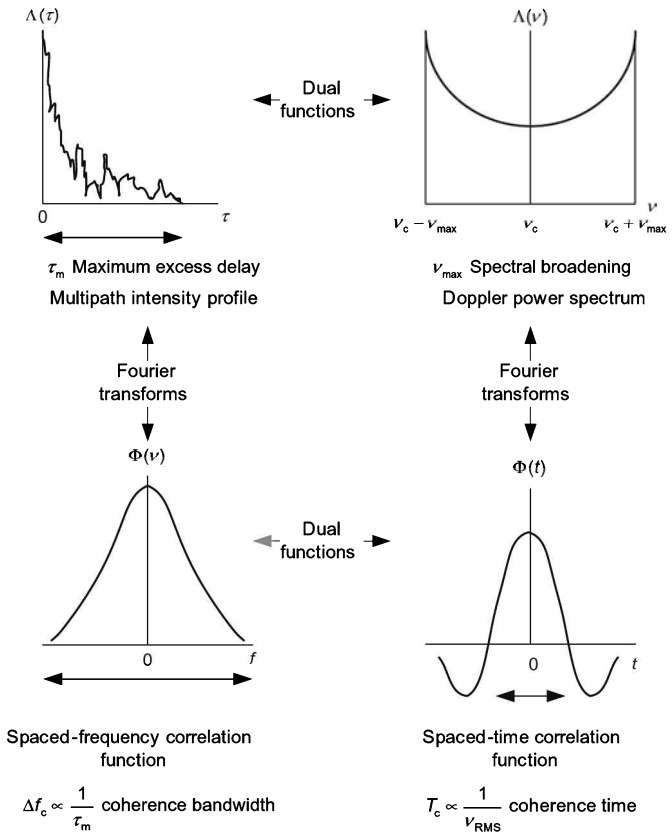


FIGURE 9.3

Illustration of the duality principle in time- and frequency-domains [19,48,49].

continuum of multipath components in some channels such as the tropospheric channel. In order to perform measurements of the multipath intensity profile, a wideband signal, i.e., a unit impulse or Dirac delta function, is used. For a single transmitted impulse, the time τ_m between the first and last received component is defined as the maximum excess delay during which the multipath signal power typically falls to some level 10–20 dB below that of the strongest component. Note that for an ideal system with zero excess delay, the function $\Lambda(\tau)$ would consist of an ideal impulse with weight equal to the total average received signal power. In the literature, the Fourier transform of $\Lambda(\tau)$ is referred to as the spaced-frequency correlation function $\Phi(\upsilon)$ [19]. The spaced-frequency correlation function $\Phi(\upsilon)$ is the channel's response to a pair of sinusoidal signals separated in frequency by υ. The coherence bandwidth Δf_c is a measure of the frequency range over which spectral components have a strong likelihood of amplitude correlation. In other words, a signal's spectral components over this range are affected by the channel in a similar manner. Note that Δf_c and τ_m are inversely proportional ($\Delta f_c \propto 1/\tau_m$). The maximum excess delay τ_m is not the best indicator of how any given wireless system will perform over a communication channel, because different channels with the same value of τ_m can exhibit different variations of signal intensity over the delay span. A useful measure of delay spread is most often characterized in terms of the root mean square (RMS) delay spread τ_{RMS} where [20]:

$$\tau_{RMS} = \sqrt{\frac{\int_0^{\tau_m} (\tau - \overline{\tau})^2 \Lambda(\tau) d\tau}{\int_0^{\tau_m} \Lambda(\tau) d\tau}} \tag{9.5}$$

in which the average multipath delay is calculated as follows:

$$\overline{\tau} = \frac{\int_0^{\tau_m} \tau \Lambda(\tau) d\tau}{\int_0^{\tau_m} \Lambda(\tau) d\tau} \tag{9.6}$$

An exact relationship between coherence bandwidth and delay spread does not exist and must be derived from signal analysis of actual signal dispersion measurements in specific channels. If coherence bandwidth is defined as the frequency interval over which the channel's complex-valued frequency transfer function has a correlation of at least 0.9, the coherence bandwidth is approximately $\Delta f_c \approx 1/(50\tau_{RMS})$. A common approximation of Δf_c corresponding to a frequency range over which the channel transfer function has a correlation of at least 0.5 is $\Delta f_c \approx 1/(5\tau_{RMS})$.

A channel is said to exhibit frequency-selective effects if $\Delta f_c < 1/\tau_s$, where the inverse symbol rate is approximately equal to the signal bandwidth W. In

practice, W may differ from $1/\tau_s$ due to filtering or data modulation. Frequency-selective fading effects arise whenever a signal's spectral components are not affected equally by the channel. This occurs whenever $\Delta f_c < W$. Frequency non-selective or flat fading degradation occurs whenever $\Delta f_c > W$. Hence, all of the signal's spectral components will be affected by the channel in a similar manner. Flat-fading does not introduce channel-induced ISI distortion, but performance degradation can still be expected due to loss in SNR whenever the signal experiences fading. In order to avoid channel-induced ISI distortion, the channel is required to exhibit flat-fading by ensuring that $\Delta f_c > W$. Therefore, the channel coherence bandwidth Δf_c sets an upper limit on the transmission rate that can be used without incorporating an equalizer in the receiver.

Figure 9.3 shows function $\Phi(t)$ which is known as the spaced-time correlation function, which is the autocorrelation function of the channel's response to a sinusoid. This function specifies the extent to which there is correlation between the channel's response to a sinusoid sent at time t_1 and the response to a similar sinusoid sent at time t_2, where $\Delta t = t_2 - t_1$. The coherence time is a measure of the expected time duration over which the channel's response is essentially invariant. To estimate $\Phi(t)$, a sinusoidal signal is transmitted through the channel and the autocorrelation function of the channel output is calculated. The function $\Phi(t)$ and the parameter T_c provide information about the speed of fading-channel variation. Note that for an ideal time-invariant channel, the channel's response would be highly correlated for all values of Δt and $\Phi(t)$ would be a constant function. If one ideally assumes uniformly distributed scattering around a mobile station with linearly polarized antennas, then the Doppler power spectrum (i.e., the inverse Fourier transform of spaced-time correlation function $\Lambda(\nu)$) has a U-shaped distribution, as shown in Figure 9.3. In a time-varying fading channel, the channel response to a pure sinusoidal tone spreads over a finite frequency range $\nu_c - \nu_{max} < \nu < \nu_c + \nu_{max}$, where ν_c and ν_{max} denote the frequency of the sinusoidal tone and the maximum Doppler spread, respectively. The RMS bandwidth of $\Lambda(\nu)$ is referred to as Doppler spread and is denoted by ν_{RMS} that can be calculated as follows:

$$\nu_{RMS} = \sqrt{\frac{\int_{\nu_c-\nu_{max}}^{\nu_c+\nu_{max}} (\nu-\overline{\nu})^2 \Lambda(\nu)\mathrm{d}\nu}{\int_{\nu_c-\nu_{max}}^{\nu_c+\nu_{max}} \Lambda(\nu)\mathrm{d}\nu}} \tag{9.7}$$

where:

$$\overline{\nu} = \frac{\int_{\nu_c-\nu_{max}}^{\nu_c+\nu_{max}} \nu\Lambda(\nu)\mathrm{d}\nu}{\int_{\nu_c-\nu_{max}}^{\nu_c+\nu_{max}} \Lambda(\nu)\mathrm{d}\nu} \tag{9.8}$$

Samsung Ex. 1005, Page 439 of 1147

The coherence time is typically defined as the time lag for which the signal autocorrelation coefficient reduces to 0.7. The coherence time is inversely proportional to Doppler spread $T_c \approx 1/\nu_{RMS}$. A common approximation for the value of coherence time as a function of Doppler spread is $T_c = 0.423/\nu_{RMS}$. It can be observed that the functions on the right-hand side of Figure 9.3 are dual of the functions on the left-hand side (duality principle).

The angle of spread refers to the spread in angle of arrival (AoA) of the multipath components at the receiver antenna array. At the transmitter, on the other hand, the angle of spread refers to the spread in the angle of departure (AoD) of the multipath components that leave the transmit antennas. If the angle spectrum function $\Theta(\theta)$ denotes the average power as a function of AoA, then the RMS angle spread can be calculated as follows:

$$\theta_{RMS} = \sqrt{\frac{\int_{-\pi}^{+\pi} (\theta - \bar{\theta})^2 \Theta(\theta) d\theta}{\int_{-\pi}^{+\pi} \Theta(\theta) d\theta}} \tag{9.9}$$

where:

$$\bar{\theta} = \frac{\int_{-\pi}^{+\pi} \theta \Theta(\theta) d\theta}{\int_{-\pi}^{+\pi} \Theta(\theta) d\theta} \tag{9.10}$$

The angle of spread causes space-selective fading, which manifests itself as variation of signal amplitude according to the location of antennas. The space-selective fading is characterized by the coherence distance D_c which is the spatial separation for which the autocorrelation coefficient of the spatial fading reduces to 0.7. The coherence distance is inversely proportional to the angle of spread $D_c \propto 1/\theta_{RMS}$.

In Figure 9.3, a duality between multipath intensity function $\Lambda(\tau)$ and Doppler power spectrum $\Lambda(\nu)$ is identified. It means that the two functions exhibit similar behavior across the time-domain and frequency-domain. As the $\Lambda(\tau)$ function identifies expected power of the received signal as a function of delay, $\Lambda(\nu)$ identifies expected power of the received signal as a function of frequency. Similarly, the spaced-frequency correlation function $\Phi(f)$ and spaced-time correlation function $\Phi(t)$ are dual functions. It implies that $\Phi(f)$ represents channel correlation in frequency-domain and $\Phi(t)$ corresponds to channel correlation function in time-domain in a similar manner.

Having reviewed some underlying concepts and characteristics of the wireless channels, let us now focus on some practical aspects of communication (data transmission) over wireless channels, and in particular the use of the above properties in the formation of a channel vector/matrix. As mentioned earlier, a wireless channel exhibits time, frequency, and space-selective variations known as fading. This fading arises due to Doppler, delay, and angle spreads in the scattering environment. The

Samsung Ex. 1005, Page 440 of 1147

channel spreading can be observed by sending a single impulse in frequency or continuous-wave signal in time or angle (point source) through the channel and receiving a signal spread along the spectral, temporal, or spatial dimension, respectively. A frequency-flat solution, however, can be applied to a frequency-selective channel by decomposing the transmission band into multiple narrow, frequency-flat sub-bands. Specifically, we can apply the solution per sub-carrier in systems employing OFDM. In a rich scattering environment, a frequency-flat MIMO wireless channel can be modeled as a complex Gaussian random process, represented as a time-varying matrix. The channel at each time instance is a Gaussian random variable, specified by the mean and its covariance. A non-zero channel mean signifies the presence of a direct LoS, or a cluster of strong paths, and the channel envelop has the Rician statistics, while zero mean corresponds to the Rayleigh statistics. The channel covariance, on the other hand, encompasses the correlation among the antennas at the transmitter and the receiver. Assuming the channel is stationary, the channel temporal variation can be captured by the channel auto-covariance, measuring the correlation between two channel instances separated by a delay. At zero delay, the channel auto-covariance coincides with the channel covariance [129].

Considering the channel state information at the transmitter (CSIT) at the transmit time in the form of a channel estimate and the estimation error covariance derived from a channel measurement at an initial time and the channel statistics, the main source of irreducible error in channel estimation is the random time variation of the channel between the initial measurement and its use by the transmitter. Therefore, we assume that the initial channel measurement is error free. The error in the channel estimate depends only on the time delay and the channel time selectivity, or the Doppler spread.

Let $\mathbf{H}$ denote the $N_{\mathrm{RX}} \times N_{\mathrm{TX}}$ channel matrix in a system with N_{TX} transmit and N_{RX} receive antennas. The channel has statistical mean $\overline{\mathbf{H}} = \mathscr{E}[\mathbf{H}]$ and covariance $\mathbf{R}_{\mathbf{H}} = \mathscr{E}[\mathbf{h}\mathbf{h}^H] - \overline{\mathbf{h}}\,\overline{\mathbf{h}}^H$, where the lower-case letter denotes the vectorized version of the upper-case matrix. Let us further assume that we have an initial, accurate channel measurement $\mathbf{H}(0)$. The channel auto-covariance $\mathbf{R}(\tau)$ at time delay τ then indicates the correlation between the initial measurement $\mathbf{H}(0)$ and the current channel $\mathbf{H}(\tau)$, defined as $\mathbf{R}(\tau) = \mathscr{E}[\mathbf{h}(\tau)\mathbf{h}^H(0)] - \overline{\mathbf{h}}(\tau)\overline{\mathbf{h}}^H(0)$. Intuitively, when this correlation is strong (i.e., $\mathbf{R}(\tau)$ is large when measured in a suitable norm) then $\mathbf{H}(0)$ is useful for estimating $\mathbf{H}(\tau)$. The strongest correlation is when the delay is zero (i.e., if $\tau \to 0$, then $\mathbf{R}(\tau) \to \mathbf{R}(0)$). In a scalar system, $\mathbf{R}(\tau)$ and $\mathbf{R}(0)$ reduce to scalars $r(\tau)$ and $r(0)$, respectively, and they are related through $r(\tau) = \rho(\tau)r(0)$, where $|\rho(\tau)| \leq 1$ is the temporal correlation coefficient [129].

An important assumption about channel temporal homogeneity is that the temporal correlation-coefficient $\rho(\tau)$ between any pair of transmit-antenna and receive-antenna is identical. This assumption is based on the premise that the channel temporal statistics can be expected to be the same for all antenna pairs. It is now possible to separate the temporal correlation from the spatial correlation in the channel auto-covariance as $\mathbf{R}(\tau) = \rho(\tau)\mathbf{R}(0)$. The temporal correlation $\rho(\tau)$ is a function of the time delay τ and the channel Doppler spread. As an example, in

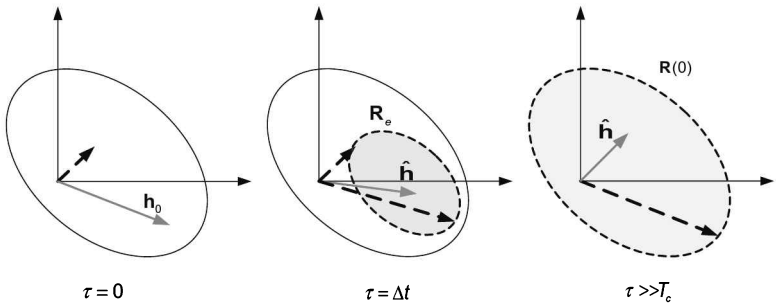


FIGURE 9.4

Illustration of a dynamic CSIT model [129].

Jake's model [21], $\rho(\tau) = J_0(2\pi\tau\nu)$, where ν is the channel Doppler spread and $J_0(\cdot)$ is the zero-order Bessel function of the first kind [35–37].

It can be shown that an estimate of the channel at time τ together with the estimation error covariance using minimum mean-squared error (MMSE) criterion results in the following $\hat{\mathbf{H}}(\tau) = \rho\mathbf{H}(0) - (1-\rho)\overline{\mathbf{H}}$ and $\mathbf{R}_e(\tau) = (1-\rho^2)\mathbf{R}(0)$. The two quantities $\hat{\mathbf{H}}(\tau)$ and $\mathbf{R}_e(\tau)$ function effectively as a new channel mean and a new channel covariance; thus they are referred to as the effective mean and the effective covariance, respectively, and constitute the CSIT. This CSIT ranges from perfect channel knowledge when $\rho = 1$, to pure stochastic when $\rho = 0$. Since the CSIT depends on the value of ρ which captures the channel time variation, it is called dynamic CSIT. The correlation factor ρ functions as a measure of CSIT quality. When $\rho = 1$, the channel estimate coincides with $\mathbf{H}(0)$ and is error free. As ρ approaches to zero, the effect of the initial channel measurement diminishes and the estimate asymptotically approaches the channel mean $\overline{\mathbf{H}}$. In parallel, the estimation error covariance $\mathbf{R}_e$ is zero when $\rho = 1$, and approaches zero as the correlation factor approaches zero. Figure 9.4 illustrates the CSIT variation as a function of the time delay τ. Several special cases of dynamic CSIT are of interest as follows: (1) perfect CSIT, in which the effective covariance is zero and the effective mean is the instantaneous channel matrix; (2) mean CSIT, in which the effective mean is non-zero and arbitrary, but the effective covariance is the identity matrix, corresponding to uncorrelated antennas; (3) covariance CSIT, in which the effective covariance matrix is non-identity and arbitrary, but the effective mean is zero, corresponding to Rayleigh fading. The general case in which both the mean and covariance matrices are arbitrary is referred to as statistical CSIT [129].

9.3 LTE/LTE-Advanced operating frequencies and band classes

The LTE/LTE-Advanced were designed to operate in the frequency bands defined in Table 9.1. The requirements were defined for 1.4, 3, 5, 10, 15, and 20 MHz

Table 9.1 E-UTRA Operating Frequencies, Band Classes, and Supported Bandwidths [1,2]

E-UTRA Operating Band (Band Class)	Uplink Operating Band eNB Receive/UE Transmit (MHz) $f_{UL\text{-}low} - f_{UL\text{-}high}$	$N_{Offset\text{-}DL}$	N_{DL}	Downlink Operating Band eNB Transmit/UE Receive (MHz) $f_{DL\text{-}low} - f_{DL\text{-}high}$	$N_{Offset\text{-}UL}$	N_{UL}	Duplex Scheme	Supported Bandwidths (MHz
1	1920–1980	0	0–599	2110–2170	18000	18000–18599	FDD	5, 10, 15, 20
2	1850–1910	600	600–1199	1930–1990	18600	18600–19199	FDD	1.4, 3, 5, 10, 15, 20
3	1710–1785	1200	1200–1949	1805–1880	19200	19200–19949	FDD	1.4, 3, 5, 10, 15, 20
4	1710–1755	1950	1950–2399	2110–2155	19950	19950–20399	FDD	1.4, 3, 5, 10, 15, 20
5	824–849	2400	2400–2649	869–894	20400	20400–20649	FDD	1.4, 3, 5, 10
6	830–840	2650	2650–2749	875–885	20650	20650–20749	FDD	5, 10
7	2500–2570	2750	2750–3449	2620–2690	20750	20750–21449	FDD	5, 10
8	880–915	3450	3450–3799	925–960	21450	21450–21799	FDD	1.4, 3, 5, 10
9	1749.9–1784.9	3800	3800–4149	1844.9–1879.9	21800	21800–22149	FDD	5, 10, 15, 20
10	1710–1770	4150	4150–4749	2110–2170	22150	22150–22749	FDD	5, 10, 15, 20
11	1427.9–1447.9	4750	4750–4949	1475.9–1495.9	22750	22750–22949	FDD	5, 10
12	699–716	5010	5010–5179	729–746	23010	23010–23179	FDD	1.4, 3, 5, 10
13	777–787	5180	5180–5279	746–756	23180	23180–23279	FDD	5, 10
14	788–798	5280	5280–5379	758–768	23280	23280–23379	FDD	5, 10
15	Reserved			Reserved			FDD	–
16	Reserved			Reserved			FDD	–
17	704–716	5730	5730–5849	734–746	23730	23730–23849	FDD	5, 10
18	815–830	5850	5850–5999	860–875	23850	23850–23999	FDD	5, 10, 15
19	830–845	6000	6000–6149	875–890	24000	24000–24149	FDD	5, 10, 15
20	832–862	6150	6150–6449	791–821	24150	24150–24449	FDD	5, 10, 15, 20

21	1447.9–1462.9	6450	6450–6599	1495.9–1510.9	24450	24450–24599	FDD	5, 10, 15
22	3410–3490	6600	6600–7399	3510–3590	24600	24600–25399	FDD	5, 10, 15, 20
23	2000–2020	7500	7500–7699	2180–2200	25500	25500–25699	FDD	1.4, 3, 5, 10
24	1626.5–1660.5	7700	7700–8039	1525–1559	25700	25700–26039	FDD	5, 10
25	1850–1915	8040	8040–8689	1930–1995	26040	26040–26689	FDD	1.4, 3, 5, 10, 15, 2
...				Not Specified				
33	1900–1920	36000	36000–36199	1900–1920	36000	36000–36199	TDD	5, 10, 15, 20
34	2010–2025	36200	36200–36349	2010–2025	36200	36200–36349	TDD	5, 10, 15
35	1850–1910	36350	36350–36949	1850–1910	36350	36350–36949	TDD	1.4, 3, 5, 10, 15, 2
36	1930–1990	36950	36950–37549	1930–1990	36950	36950–37549	TDD	1.4, 3, 5, 10, 15, 2
37	1910–1930	37550	37550–37749	1910–1930	37550	37550–37749	TDD	5, 10, 15, 20
38	2570–2620	37750	37750–38249	2570–2620	37750	37750–38249	TDD	5, 10, 15, 20
39	1880–1920	38250	38250–38649	1880–1920	38250	38250–38649	TDD	5, 10, 15, 20
40	2300–2400	38650	38650–39649	2300–2400	38650	38650–39649	TDD	5, 10, 15, 20
41	2496–2690	39650	39650 –41589	2496–2690	39650	39650 –41589	TDD	5, 10, 15, 20
42	3400–3600	41590	41590–43589	3400–3600	41590	41590–43589	TDD	5, 10, 15, 20
43	3600–3800	43590	43590–45589	3600–3800	43590	43590–45589	TDD	5, 10, 15, 20

bandwidth with a specific configuration in terms of the number of physical resource blocks (PRBs). Using carrier aggregation, a number of contiguous and/or non-contiguous frequency bands can be aggregated to create a virtually larger bandwidth. The channel raster is 100 kHz, which means the center frequency must be a multiple of 100 kHz [1]. To support transmission in paired and unpaired spectra, two duplexing schemes are supported, i.e., frequency division duplex (FDD), allowing both full and half-duplex terminal operation, as well as time division duplex (TDD). Table 9.2 gives the band classes where the LTE/LTE-Advanced systems can be deployed.

Figure 9.5 illustrates the relation between the channel bandwidth (BW_{Channel}) and the transmission bandwidth, i.e., the number of permissible PRBs (N_{RB}). The channel edges are defined as the lowest and highest frequencies of the carrier separated by the channel bandwidth; thus $f_C \pm (1/2)BW_{\text{Channel}}$. In the case of carrier aggregation, the aggregated channel bandwidth $BW_{\text{Channel-CA}}$ is defined as $BW_{\text{Channel-CA}} = f_{\text{edge-high}} - f_{\text{edge-low}}$. The lower band edge $f_{\text{edge-low}}$ and the upper band edge $f_{\text{edge-high}}$ of the aggregated channel bandwidth are used as the reference points in frequency for transmitter and receiver requirements, and are subsequently defined as follows: $f_{\text{edge-low}} = f_{\text{C-low}} - f_{\text{offset-low}}$ and $f_{\text{edge-high}} = f_{\text{C-high}} + f_{\text{offset-high}}$. The lower and upper frequency offsets depend on the transmission bandwidth configurations of the lowest and highest assigned edge component carrier, and are defined as $f_{\text{offset-low}} = 0.18N_{\text{RB-low}}/2 + BW_{\text{Guard-band}}$ and $f_{\text{offset-high}} = 0.18N_{\text{RB-high}}/2 + BW_{\text{Guard-band}}$, where $N_{\text{RB-low}}$ and $N_{\text{RB-high}}$ denote the transmission bandwidth configurations for the lowest and highest assigned component carrier, respectively, and $BW_{\text{Guard-band}}$ denotes the nominal guard band. Note that the 0.18 multiplier in the equations is the bandwidth of a PRB in MHz. The aggregated transmission bandwidth is defined as the number of the aggregated resource blocks within the entire assigned aggregated channel bandwidth, and is defined per carrier aggregation band class.

The frequency separation between carriers depends on the deployment scenario and the size of the frequency block available to the network operator, as well as the channel bandwidth. The nominal channel spacing between two adjacent E-UTRA carriers is defined as $\Delta f_{\text{spacing}} = (BW_{\text{Channel(1)}} + BW_{\text{Channel(2)}})/2$, where $BW_{\text{Channel(1)}}$ and $BW_{\text{Channel(2)}}$ parameters are the channel bandwidths of the two respective E-UTRA carriers. The channel spacing can be adjusted to optimize performance in a particular deployment scenario. For intra-band contiguous carrier aggregation bandwidth, the nominal channel spacing between two adjacent E-UTRA component carriers is specified as follows:

$$\Delta f_{\text{spacing-CA}} = 0.3 \left\lfloor \frac{BW_{\text{Channel(1)}} + BW_{\text{Channel(2)}} - 0.1\left|BW_{\text{Channel(1)}} - BW_{\text{Channel(2)}}\right|}{0.6} \right\rfloor \text{(MHz)} \tag{9.11}$$

The channel spacing for intra-band contiguous carrier aggregation can be adjusted to any integer multiple of 300 kHz (less than the nominal channel spacing)

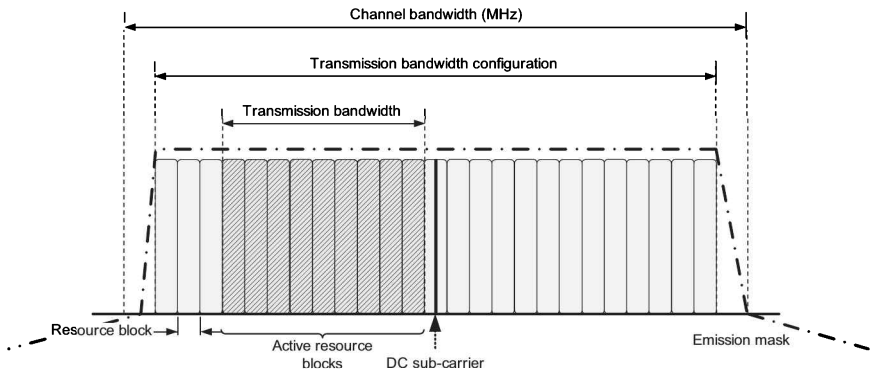


FIGURE 9.5

Relationship between channel bandwidth and transmission bandwidth in E-UTRA [1].

to optimize performance in a particular deployment scenario. The E-UTRA channel raster (including the carrier aggregation scenarios) is 100 kHz for all bands, which means that the carrier center frequency must be an integer multiple of 100 kHz.

The carrier frequency in the downlink/uplink directions is designated by the E-UTRA absolute radio frequency channel number (EARFCN) in the range of 0–65,535. The relation between EARFCN and the downlink carrier frequency (in MHz) is given by $f_{DL} = f_{DL-low} + 0.1(N_{DL} - N_{Offset-DL})$, where f_{DL-low}, N_{DL}, and $N_{Offset-DL}$ parameters are given in Table 9.1. Similarly, the relationship between EARFCN and the uplink carrier frequency (in MHz) for the uplink is defined as $f_{UL} = f_{UL-low} + 0.1(N_{UL} - N_{Offset-UL})$, where f_{UL-low}, N_{UL}, and $N_{Offset-UL}$ parameters are provided in Table 9.1. Note that N_{DL} and N_{UL} in the latter equations are the downlink and uplink EARFCN, respectively.

The channel numbers that designate carrier frequencies which are extremely close to the operating band edges are not used. This implies that the first 7, 15, 25, 50, 75, and 100 channel numbers at the lower operating band edge and the last 6, 14, 24, 49, 74, and 99 channel numbers at the upper operating band edge are not utilized for channel bandwidths of 1.4, 3, 5, 10, 15, and 20 MHz, respectively.

The fragmentation of the suitable and permissible frequency bands for deployment of LTE systems across the globe and the disparity of the operators' network deployments and legacy systems have made significant challenges against worldwide proliferation and accessibility of 4G networks (see Figure 9.6). Figure 9.6 depicts the distribution of the useable frequency bands for commercial LTE/LTE-Advanced deployments across the globe as of August 2012 [161]. In general, typical user equipment has to support multiple communication RF bands and standards (e.g., LTE/LTE-Advanced, UMTS, and 1xEV-DO) in order to allow operation across different regions and operator networks. Enabling the mobile terminals to cover all desired RF

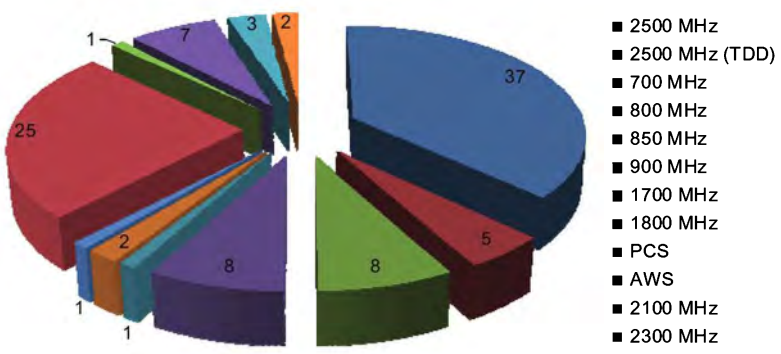


FIGURE 9.6

Distribution of frequency bands for E-UTRA commercial deployments across the world [161].

bands does not only result in an increased chip area due to more complex RF integrated circuit implementation, but also a larger phone form factor to accommodate external filters (see Figure 9.7 as an example of triple-band LTE transceiver). In the case of FDD systems, common handhelds use bulk acoustic wave duplex filters[1] to separate the receive signal from the transmit signal with a nominal frequency separation. These filters are not tunable, and thus require one duplexer per frequency constellation. The selectivity provided by the duplexer must reach a certain level such that third order inter-modulation frequency components are suppressed. These signals are generated by the transmit signal leaking into the receiver and an external interferer in a certain frequency band, thus degrading sensitivity of the entire receiver. Therefore, it is necessary to use either very expensive duplexers that could provide the desired selectivity, or to place an external (i.e., external to RFIC) low-noise amplifier followed by an additional surface acoustic wave filter[2] which is a more common solution. In general, it is not trivial to design extremely broadband RF filters, power or low-noise amplifiers, mixers, etc., to cover a number of contiguous and/or non-contiguous RF bands with optimal performance across the entire operating bands. Figure 9.8 shows the measured and simulated performance of a triple-fed

[1]Bulk acoustic wave filters are electro-mechanical devices typically operating in the frequency range of 2–16 GHz which can be implemented by ladder or lattice filters. Two main variants of bulk acoustic wave filter are thin film bulk acoustic resonator, or FBAR, and solid mounted bulk acoustic resonators [167].

[2]Surface acoustic wave filters are electro-mechanical devices commonly used in RF systems. The electrical signal is converted to a mechanical wave in a device constructed of a piezoelectric crystal or ceramic. This wave is delayed as it propagates across the device, before being converted back to an electrical signal by further electrodes. The delayed outputs are recombined to produce a direct analog implementation of a finite impulse response filter. This hybrid filtering technique has found applications in analog sampled filters. The surface acoustic wave filters are limited to frequencies up to 3 GHz [167].

Samsung Ex. 1005, Page 447 of 1147

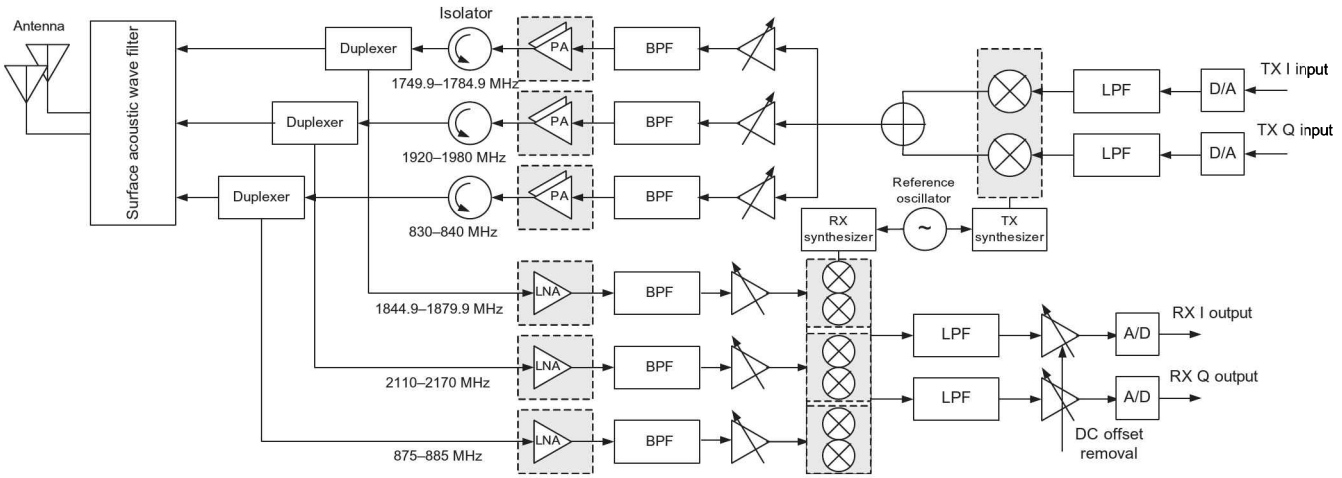


FIGURE 9.7

Example of a Triple-band LTE Transceiver [170].

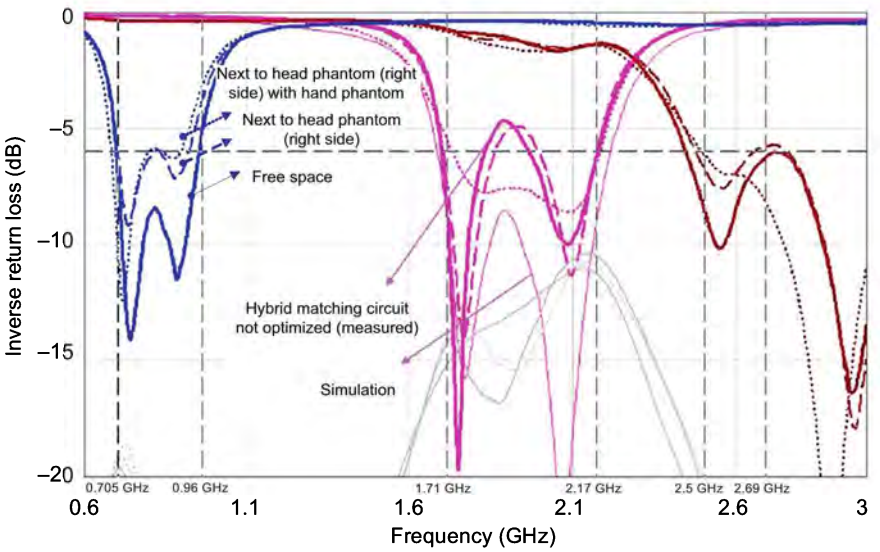


FIGURE 9.8

Simulated and measured inverse return loss of an example triple-fed antenna at 0.8, 1.9, 2.6 GHz [170].

antenna in terms of input return loss[3] (i.e., an indication of the quality of input impedance matching) under three different practical scenarios and in different frequency bands. It can be seen that satisfactory impedance matching is only achieved in a certain frequency region and not across the entire frequency bands, especially if the antennas are covered by hand or head phantoms. This example underlines the practical challenge of designing extremely wideband RF systems to fit into very small user-equipment's form factor in order to enable multi-band multi-mode LTE operation [170].

[3]Return loss is the loss of signal power resulting from the reflection caused at a discontinuity in a transmission media. This discontinuity can be a mismatch with the terminating load or with a device inserted in the line. It is usually expressed as the ratio of the incident power P_i and the reflected power P_r, i.e., RL (dB) = 10 log(P_i/P_r) in dB. The return loss is related to both the standing wave ratio (SWR) and reflection coefficient (Γ). Increasing the value of return loss corresponds to lower SWR. Return loss is a measure of how well the input/output impedances of devices or transmission lines are matched. A match is good, if the return loss value is high. A high return loss is desirable and results in a lower insertion loss. The ratio of the amplitude of the reflected wave V_r to the amplitude of the incident wave V_i is known as the reflection coefficient Γ, which is defined as $\Gamma = V_r/V_i$. When the source and load impedances are known values, the reflection coefficient is given by $\Gamma = (Z_L - Z_0)/(Z_L + Z_0)$ where Z_0 is the impedance of the transmission line toward the source and Z_L is the impedance toward the load. The return loss is the negative of the magnitude of the reflection coefficient in dB. Since power is proportional to the square of the voltage, the return loss is given by RL(dB) = $-$ 20 log(|Γ|) [169].

Samsung Ex. 1005, Page 449 of 1147

The performance of mobile device antennas is an important factor in the design and implementation of the device (in terms of practical form factor) as well as the ultimate communication quality that a user may expect from that device. This requires adopting an antenna design that yields high performance under realistic conditions in order to ensure satisfaction of LTE/LTE-Advanced system requirements. This challenge will be further intensified in the case of multi-band antenna array design which is required to fit into the small form factor of a mobile device and to provide satisfactory performance (near-field/far-field radiation patterns, gain, return loss, parasitic/spurious effects, etc.) over a large bandwidth.

9.4 Principles of OFDM

The orthogonal frequency division multiplexing (OFDM) is a form of multi-carrier modulation technique that was first introduced over 50 years ago. In a conventional serial data transmission system, the information bearing symbols are transmitted sequentially, with the frequency spectrum of each symbol occupying the entire available bandwidth. Figure 9.9 illustrates an unfiltered quadrature amplitude modulation (QAM) signal spectrum. It is in the form of $\sin(\pi f T_u)/\pi f T_u$ with zero crossing points at integer multiples of $1/T_u$, where T_u is the QAM symbol period. The concept of OFDM is to transmit the data bits in parallel QAM modulated sub-carriers using frequency division multiplexing. The carrier spacing is carefully selected so that each sub-carrier is located on other sub-carriers' zero-crossing points in the frequency-domain. Although there are spectral overlaps among sub-carriers, they do not interfere with each other if they are sampled at the sub-carrier frequencies. In other words, they maintain spectral orthogonality. As shown in Figure 9.9, the OFDM signal in the frequency-domain is generated through aggregation of N_{FFT} parallel QAM-modulated sub-carriers, where adjacent sub-carriers, are separated by sub-carrier spacing $1/T_u$. An example of OFDM signal spectrum as seen on a vector signal analyzer is shown in Figure 9.10.

Since an OFDM signal consists of many parallel QAM sub-carriers, the mathematical expression of the signal in the time-domain can be expressed as follows:

$$s(t) = \mathrm{Re}\left\{ e^{j\omega_c t} \sum_{k=-(N_{FFT}-1)/2}^{(N_{FFT}-1)/2} \alpha_k \, e^{j2\pi k(t-t_g)/T_u} \right\} \quad mT_u \le t \le (m+1)T_u \qquad (9.12)$$

where $s(t)$ denotes the OFDM signal in the time-domain, α_k is the complex-valued data that is QAM-modulated and transmitted over sub-carrier k, N_{FFT} is the number of sub-carriers in the frequency-domain, ω_c is the RF carrier frequency, and T_g is the guard interval or the cyclic prefix (CP) length. For a large number of sub-carriers, direct generation and demodulation of the OFDM signal would require arrays of coherent sinusoidal generators which can become excessively complex and expensive. However, one can notice that the OFDM signal is actually the real part of the inverse discrete Fourier transform (IDFT) of the original complex-valued data

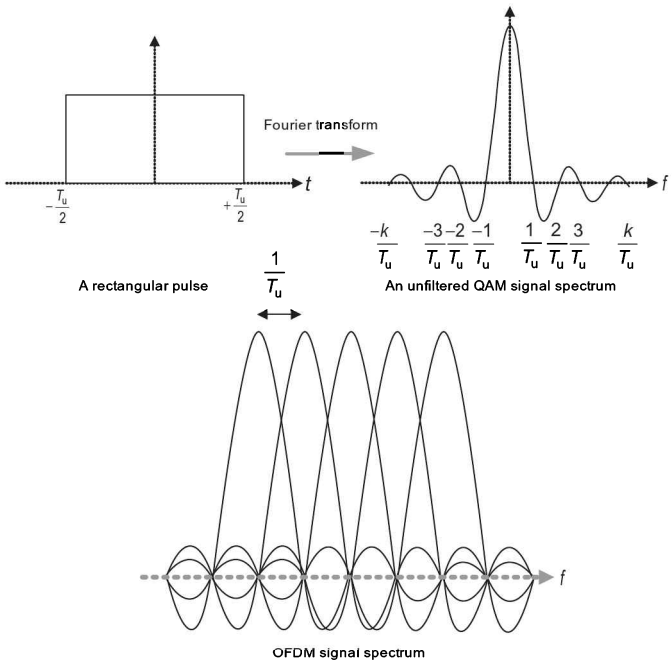


FIGURE 9.9

Illustration of OFDM concept [52].

symbols $\{\alpha_k\}|_{k=-(N_{\text{FFT}}-1)/2,\ldots,(N_{FFT}-1)/2}$. It can be seen that there are $N < N_{\text{FFT}}$ sub-carriers, each carrying the corresponding data α_k. The inverse of the sub-carrier spacing $\Delta f = 1/T_{\text{u}}$ is defined as the OFDM useful symbol duration T_{u}, which is N_{FFT} times longer than that of the original input data symbol duration.

Since IDFT is used in the OFDM modulator, the original data is defined in the frequency-domain, while the OFDM signal $s(t)$ is defined in the time-domain. The IDFT can be implemented via a computationally efficient fast Fourier transform (FFT) algorithm. The orthogonality of sub-carriers in OFDM can be maintained and individual sub-carriers can be completely separated and demodulated by an FFT at the receiver when there is no ISI introduced by the communication channel. In practice, linear distortions such as multipath delay cause ISI between OFDM symbols, resulting in loss of orthogonality and an effect that is similar to co-channel interference. However, when delay spread is small, i.e., within a fraction of the OFDM useful symbol length, the impact of ISI is insignificant, although it depends on the order of modulation implemented by the sub-carriers (see Figure 9.11). A simple solution to mitigate multipath delay is to increase the OFDM effective symbol duration such that it is much larger than the delay spread; however, when the delay spread is large, it requires a large number of sub-carriers and a large FFT size. Meanwhile, the system might become sensitive to Doppler shift and carrier frequency offset (CFO). An alternative approach to mitigate

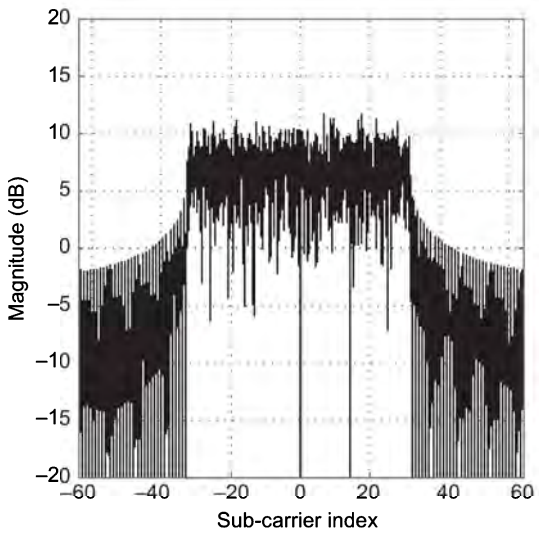


FIGURE 9.10

A snapshot of OFDM signal in the frequency-domain [158].

multipath distortion is to generate a cyclically extended guard interval, where each OFDM symbol is prefixed with a periodic extension of the signal itself, as shown in Figure 9.12 where the tail of the symbol is copied to the beginning of the symbol. The OFDM symbol duration is then defined as $T_s = T_u + T_g$, where T_g is the guard interval or cyclic prefix. When the guard interval is longer than the channel impulse response or the multipath delay, the ISI can be effectively eliminated. The ratio of the guard interval to useful OFDM symbol duration depends on the deployment scenario and the frequency band. Since the insertion of the guard intervals will reduce data throughput, T_g is usually selected less than $T_u/4$. The cyclic prefix should absorb most of the signal energy dispersed by the multipath channel. The entire ISI energy is contained within the cyclic prefix if its length is greater than that of the channel RMS delay spread ($T_g > \tau_{RMS}$). In general, it is sufficient to have most of the delay spread energy absorbed by the guard interval, considering the inherent robustness of large OFDM symbols to time dispersion.

The mapping of the modulated data symbol into multiple sub-carriers also allows an increase in the symbol duration. The symbol duration obtained through an OFDM scheme is much larger than that of a single carrier modulation technique with a similar transmission bandwidth. In general, when the channel delay spread exceeds the guard time, the energy contained in the ISI will be much smaller with respect to the useful OFDM symbol energy, as long as the symbol duration is much larger than the channel delay spread, i.e., $T_s \gg \tau_{RMS}$. Although large OFDM symbol duration is desirable to mitigate the ISI effects caused by time dispersion, large OFDM symbol duration can further reduce the ability to alleviate the effects of fast fading, particularly if the symbol period is large compared to the channel coherence

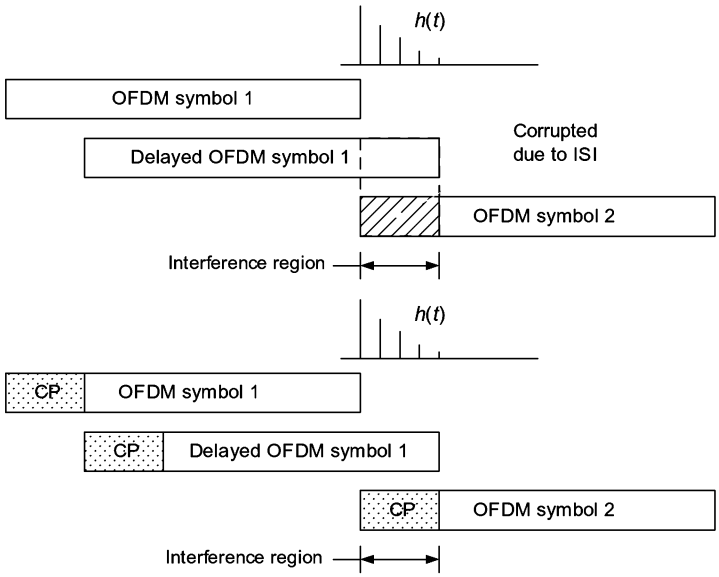


FIGURE 9.11

Illustration of the effect of guard interval for eliminating ISI ($h(t)$ is the hypothetical channel impulse response).

time, then the channel can no longer be considered as time-invariant over the OFDM symbol duration; therefore, this will introduce the inter-sub-carrier orthogonality loss. This can affect the performance in fast fading conditions. Hence, the symbol duration should be kept smaller than the minimum channel coherence time. Since the channel coherence time is inversely proportional to the maximum Doppler spread, the symbol duration T_s must, in general, be chosen such that $T_s \ll 1/\nu_{\text{RMS}}$. The large number of OFDM sub-carriers makes the bandwidth of the individual sub-carriers small relative to the overall signal bandwidth. With an adequate number of sub-carriers, the inter-carrier spacing is much smaller than the channel coherence bandwidth. Since the channel coherence bandwidth is inversely proportional to the channel delay spread τ_{RMS}, the sub-carrier separation is generally designed such that $\Delta f \ll 1/\tau_{\text{RMS}}$. In this case, the fading on each sub-carrier is flat and can be modeled as a complex-valued constant channel gain. The individual reception of the QAM symbols transmitted on each sub-carrier is therefore simplified to the case of a flat-fading channel. This enables a straightforward introduction of advanced MIMO schemes. Furthermore, in order to mitigate Doppler spread effects, the inter-carrier spacing should be much larger than the RMS Doppler spread $\Delta f \gg \nu_{\text{RMS}}$. Since the OFDM sampling frequency is typically larger than the actual signal bandwidth, only a subset of sub-carriers are used to carry QAM symbols. The remaining sub-carriers are left inactive prior to the IFFT, and are referred to as guard sub-carriers. The split between the active and the inactive

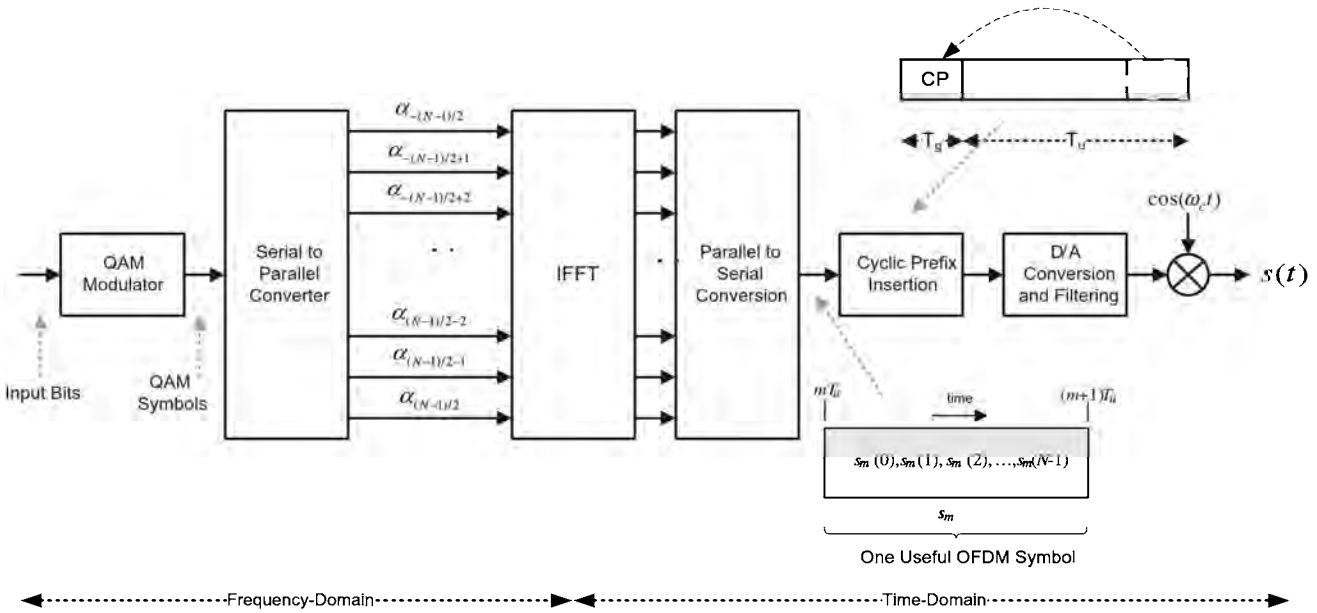


FIGURE 9.12

OFDM signal generation.

sub-carriers is determined based on the spectral sharing and regulatory constraints, such as the bandwidth allocation and the spectral mask.

An OFDM transmitter diagram is shown in Figure 9.12. The incoming bit stream is first QAM modulated to form the complex-valued QAM symbols. The QAM symbols are converted from serial to parallel with $N < N_{\text{FFT}}$ complex-valued numbers per block, where N_{FFT} is the size of FFT/IFFT operation. Each block is processed by an IFFT and the output of the IFFT forms an OFDM symbol, which is converted back to serial data for transmission. A guard interval or cyclic prefix is inserted between symbols to eliminate ISI effects caused by multipath distortion. The discrete symbols are filtered and converted to analog for RF up-conversion.

The reverse process is performed at the receiver. A one-tap equalizer is usually used for each sub-carrier to correct channel distortion. The tap coefficients are calculated based on channel information. As shown in Figures 9.9 and 9.10, one can understand that the OFDM signal spectrum is approximately rectangular. Since the OFDM modulator is an IFFT processor, its role is to convert data from the frequency-domain to the time-domain, and then transmit them in the time-domain over the channel. If a spectrum analyzer, which converts the input from the time-domain to the frequency-domain, is used to monitor an OFDM signal, what will be displayed is the original data.

When there is multipath distortion, a conventional single carrier wideband transmission system suffers from frequency selective fading. A complex adaptive equalizer must be used to equalize the in-band fading. The number of taps required for the equalizer is proportional to the symbol rate and the multipath delay. For an OFDM system, if the guard interval is larger than the multipath delay, the ISI can be eliminated and orthogonality can be maintained among sub-carriers. Since each OFDM sub-carrier occupies a very narrow spectrum, in the order of a few kHz, even under severe multipath distortion, they are only subject to flat fading. In other words, the OFDM converts a wideband frequency-selective fading channel to a series of narrowband frequency non-selective fading sub-channels by using the parallel multi-carrier transmission scheme (Figure 9.13). Since OFDM data sub-carriers are statistically independent and identically distributed, based on the central limit theorem, when the number of sub-carriers N_{FFT} is large, the OFDM signal distribution tends to be Gaussian.

In an OFDM system, the SNR is a measure of channel quality and is a key factor of link-level error assessment. There are different methods for calculation of SNR in single-antenna and multi-antenna transmission systems. For single-input/single-output systems, the SNR can be viewed as the received SNR, i.e., the received SNR before the detector. The post-processing SNR is often used for MIMO links, and represents the SNR after combining in the receiver and measures the likelihood that a coded message is decoded successfully. In link-level simulations, the SNR γ is typically calculated using the following method. Let vector $\mathbf{x} = (x_1, x_2, \ldots, x_{N_{\text{TX}}})^H \in \mathbb{C}^{N_{\text{TX}} \times 1}$ denote the transmit signal, where $x_k \in \mathbb{C} \forall k = 1, 2, \ldots, N_{\text{TX}}$ is the complex-valued transmitted symbols from the kth transmit antenna. Note that N_{TX} is the number of transmit antennas. It can be shown that the total transmit signal power can be obtained as $\sigma_{\mathbf{x}}^2 = \text{trace}(\mathbf{R}_{\mathbf{x}})$,

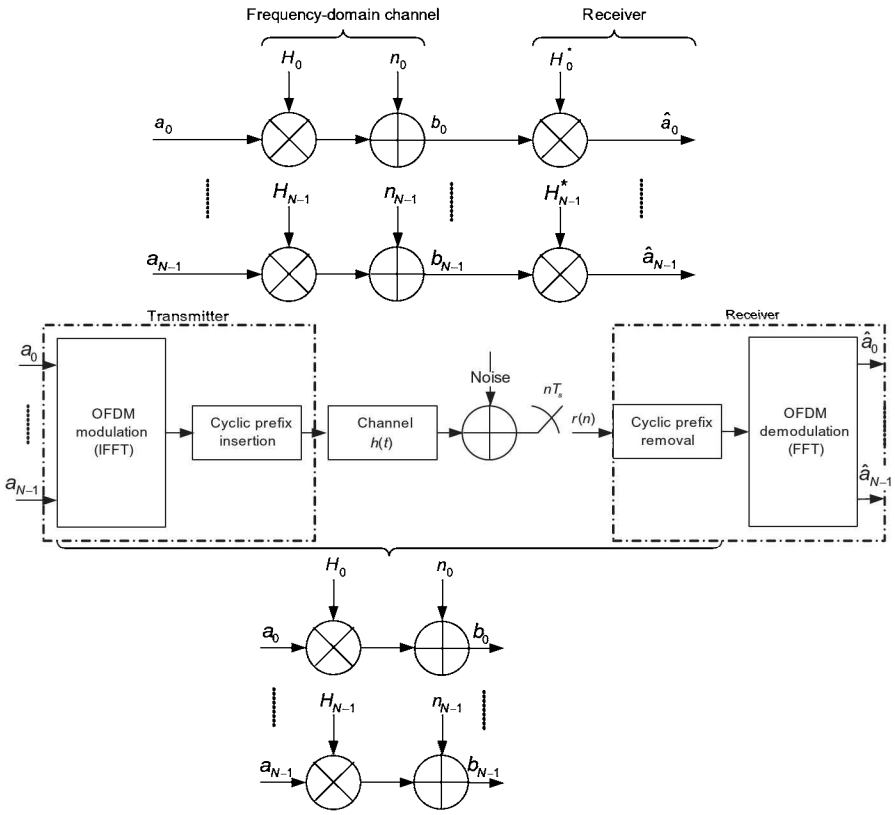


FIGURE 9.13

Single-tap equalization of the OFDM signal [14].

where $\mathbf{R_x}$ denotes the autocorrelation matrix of the transmitted signal. The transmit power from the kth antenna is given as $\sigma_{x_k}^2 = \mathscr{E}\{|x_k|^2\} = 1/N_{\text{TX}}$ (uniformly distributed power). If $\mathbf{H}$ represents the channel matrix with $||\mathbf{H}||_F^2 = N_{\text{TX}}N_{\text{RX}}$[4] in which N_{TX} and N_{RX} denote the number of transmit and receive antennas, respectively, the received signal vector $\mathbf{y}$ can be calculated as $\mathbf{y} = \mathbf{Hx} + \mathbf{v}$ where the complex-valued Gaussian-distributed noise vector $\mathbf{v} \sim \mathbb{Z}(\mathbf{0}, \sigma_{\mathbf{v}}^2\mathbf{I})$ denotes the noise vector with respect

[4] The p-norm of matrix $\mathbf{H}$ is defined as $\|\mathbf{H}\|_p = \left(\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}|h_{ij}|^p\right)^{1/p}$. In the special case when $p = 2$, the norm is called the Frobenius norm and for $p = \infty$ is called the maximum norm. The Frobenius norm or Hilbert–Schmidt norm of matrix $\mathbf{H}$ are similar, although the latter term is often reserved for the operators on a Hilbert space. In general, this norm can be defined in the following forms: $\|\mathbf{H}\|_p = \sqrt{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}|h_{ij}|^2} = \sqrt{\text{trace}(\mathbf{H}^H\mathbf{H})} = \sqrt{\sum_{i=0}^{\min(M,N)-1}\zeta_i^2}$ where $\mathbf{H}^H$ denotes the conjugate transpose of matrix $\mathbf{H}$ and ζ_i are the singular values of matrix $\mathbf{H}$. The Frobenius norm is further similar to the Euclidean norm on $\mathbb{R}^N$ and is obtained from an inner product on the space of all matrices. The Frobenius norm is sub-multiplicative and is very useful for numerical linear algebra [31,32,34].

to the size of the FFT N_{FFT} and the number of used sub-carriers before the detector N_{Used}. We define the complex-valued Gaussian-distributed noise vector $\mathbf{n} \sim \mathbb{Z}(\mathbf{0}, \sigma_{\mathbf{n}}^2\mathbf{I})$ to be the noise after the FFT operation. The receive SNR before the detector is given as $\gamma_{\text{pre-FFT}} = \frac{||\mathbf{Hx}||_F^2}{N_{\text{RX}}\sigma_{\mathbf{v}}^2} = \sigma_{\mathbf{v}}^{-2}$ whereas the SNR after the FFT operation is given as $\gamma_{\text{post-FFT}} = \frac{||\mathbf{Hx}||_F^2}{N_{\text{RX}}\sigma_{\mathbf{n}}^2} = \sigma_{\mathbf{n}}^{-2}$. It can be seen that the ratio of pre-FFT and post-FFT SNRs $\frac{\gamma_{\text{pre-FFT}}}{\gamma_{\text{post-FFT}}} = \frac{\sigma_{\mathbf{v}}^2}{\sigma_{\mathbf{n}}^2} = \frac{N_{\text{FFT}}}{N_{\text{Used}}}; N_{\text{Used}} \leq N_{\text{FFT}}$ is always positive, implying that the FFT operation suppresses the noise and enhances the SNR [142].

9.4.1 Peak-to-average power ratio of OFDM signals

The peak-to-average power ratio (PAPR) for a single-carrier modulation signal depends on its constellation and the pulse shaping filter roll-off factor. For a Gaussian distributed OFDM signal, the cumulative distribution function (CDF)[5] of PAPR for 99.0%, 99.9%, and 99.99% are approximately 8.3, 10.3, and 11.8 dB, respectively. Since the OFDM signal has a high PAPR, it could be clipped in the transmitter power amplifier (PA), because of its limited dynamic range or non-linearity. Higher output back-off is required to prevent performance degradation and inter-modulation products leaking into adjacent channels. Therefore, RF PAs should be operated in a very large linear region. Otherwise, the signal peaks leak into the non-linear region of the PA causing signal distortion. This signal distortion introduces inter-modulation among the sub-carriers and out-of-band emission. Thus, the PAs should be operated with large power back-offs. On the other hand, this leads to very inefficient amplification and expensive transmitters. Thus, it is highly desirable to reduce the PAPR. In addition to inefficient operation of the PA, a high PAPR requires a larger dynamic range for the receiver analog-to-digital (A/D) converter. To reduce the PAPR, several techniques have been proposed and used such as clipping,[6] channel coding, temporal windowing, tone reservation,[7] and tone injection. However,

[5]The CDF of the real-valued random variable X is defined as $x \rightarrow F_X(x) = P(X \leq x), \forall x \in \mathbb{R}$, where the right-hand side represents the probability that random variable X takes on a real value less than or equal to x. The CDF of X can be defined in terms of the probability density function $f(x)$ as $F(x) = \int_{-\infty}^{x} f(x)\text{d}x$. The complementary CDF (CCDF), on the other hand, is defined as $P(X > x) = 1 - F_X(x)$ [30].

[6]Since the OFDM signal has a high PAPR, it may be clipped in the transmitter power amplifier, because of its limited dynamic range or non-linearity. Higher output back-off is required to prevent BER degradation and intermodulation products spilling into adjacent channels. However, clipping of an OFDM signal has a similar effect to impulse interference, against which an OFDM system is inherently robust. Computer simulations show that for a coded OFDM system, clipping of 0.5% of the time results in a BER degradation of 0.2 dB. At 0.1% clipping, the degradation is less than 0.1 dB [52].

[7]In the tone reservation method, the transmitter and the receiver reserve a subset of tones or sub-carriers for generating PAPR reduction signals. Those reserved tones are not used for data transmission [53].

Samsung Ex. 1005, Page 457 of 1147

most of these methods are unable to achieve simultaneously a large reduction in PAPR with low complexity, low coding overhead, without performance degradation, and without transmitter-receiver symbol handshake. The PAPR ξ of the OFDM signal, defined in Eq. (9.12), is calculated as follows:

$$\xi = \left. \frac{\max |s(t)|^2}{\mathscr{E}\{|s(t)|^2\}} \right|_{mT_u \le t \le (m+1)T_u} \tag{9.13}$$

In the above equation, $\mathscr{E}\{.\}$ denotes the expectation operator, and m is an integer. From the central-limit theorem, for large values of N_{FFT}, the real and imaginary values of $s(t)$ (from Eq. (9.12) and without the $\mathrm{Re}\{\cdot\}$ operator) would have Gaussian distribution. Consequently, the amplitude of the OFDM signal has a Rayleigh distribution with zero mean and a variance of N_{FFT} times the variance of one complex sinusoid. Assuming the samples to be mutually uncorrelated, the CDF for the peak power per OFDM symbol is given by:

$$P(\xi > \gamma) = [1 - (1 - e^{-\gamma})^{N_{FFT}}] \tag{9.14}$$

From the above equation, it can be seen that a large PAPR occurs only infrequently due to relatively large values of N_{FFT} used in practice.

Figure 9.14 shows the CDF of OFDM PAPR for BPSK and QPSK modulation assuming a 40 MHz channel bandwidth and $N_{FFT} = 128$. We further applied a 90° phase rotation to the sub-carriers in the upper 20 MHz of the

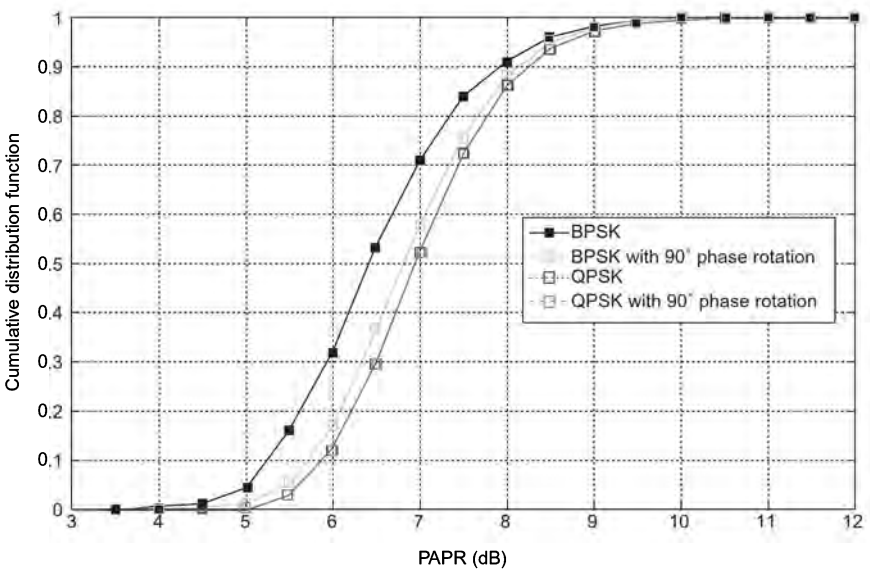


FIGURE 9.14

CDF of OFDM PAPR with BPSK/QPSK modulation and $N_{FFT} = 128$ [159].

channel, and investigated the effect on the PAPR reduction [159]. It is shown that large PAPR values are less likely to occur with large FFT sizes, as suggested by Eq. (9.14).

9.4.2 Effect of non-linearity on OFDM signals

Power amplifiers generally have a non-linear amplitude response, where the output power is saturated for large input powers. Most applications require operation in the linear region of the PA response where the output power is a linear function of the input. The larger the linear operation region, or alternatively the higher the saturation point, the more expensive the PA. Therefore, it is imperative to reduce the PAPR of the OFDM signal before processing through the PA. Figures 9.15 and 9.16 illustrate examples of PA AM/AM response and QAM signal constellation fuzziness when the OFDM signal with large PAPR is processed through a PA with a low saturation point (Figure 9.16). In this example, 64QAM modulation, 256-point IFFT with $T_{cp} = (1/4)T_u$, and PA saturation point of 6 dB are assumed.

The modulation accuracy or the permissible signal constellation fuzziness is often measured in terms of an error vector magnitude (EVM) metric. The EVM may be defined as the square root of the ratio of the mean error vector power to

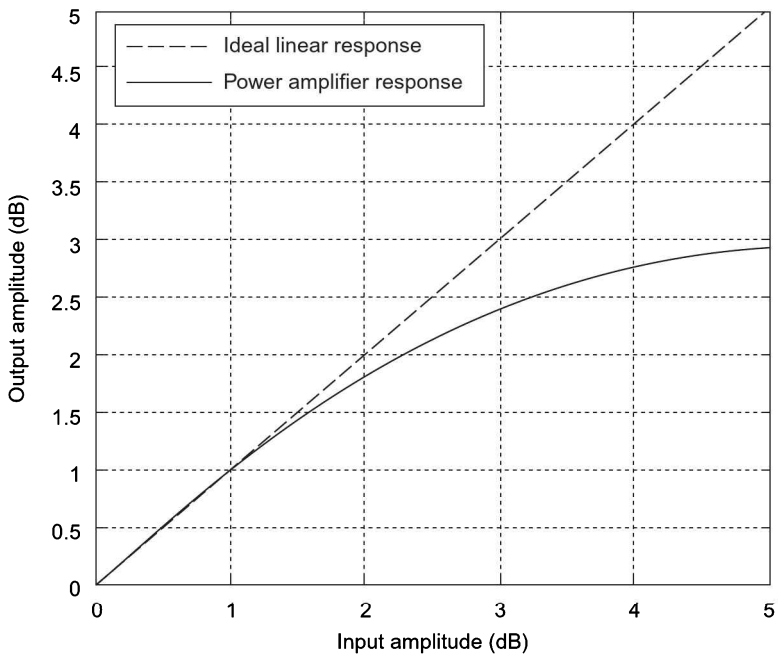


FIGURE 9.15

Example AM/AM response of a power amplifier (saturation point = 6 dB) [141].

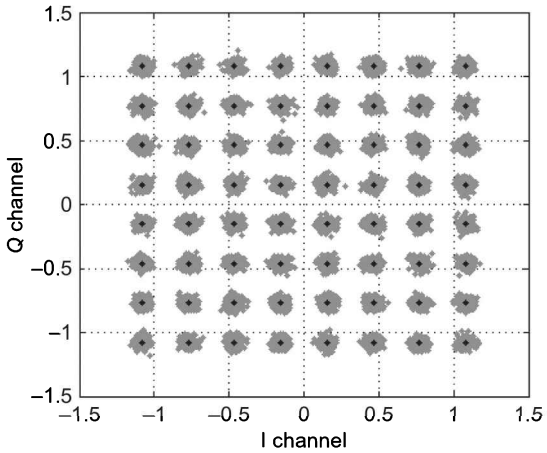


FIGURE 9.16

64QAM signal constellation before and after power amplification [141].

the mean reference signal power expressed as a percentage, i.e., the EVM can be expressed as $\mathrm{EVM} = \sqrt{\mathscr{E}|\mathbf{e}|^2/\mathscr{E}|\mathbf{s}|^2}$ in mathematical form. In other words, the EVM defines the average constellation error with respect to the farthest constellation point (i.e., the distance between the reference signal and measured signal points in the $I-Q$ plane), and is calculated as follows [1]:

$$\mathrm{EVM} = \sqrt{\frac{\sum_{i=0}^{N_{\mathrm{samples}}-1} (\Delta I_i^2 + \Delta Q_i^2)}{N_{\mathrm{samples}} r_{\mathrm{max}}^2}} \tag{9.15}$$

where N_{samples} is the number of symbols used in the measurement period, ΔI_i and ΔQ_i refer to the in-phase (I) and quadrature (Q) components of the ith error vector, and r_{max} denotes the maximum constellation amplitude. The EVM is measured over the continuous portion of a burst occupying at least one-fourth of the transmission frame at maximum power settings. The permissible EVM can be estimated from the transmitter implementation margin, if the error vector is considered noise, which is added to the channel noise. The implementation margin is the excess power needed to maintain the carrier-to-noise ratio (C/N) intact, when going from the ideal to the realistic transmitter. The EVM cannot be measured at the antenna connector, but should be measured by an ideal receiver with a certain carrier recovery loop bandwidth specified as a percentage of the symbol rate. The measured EVM includes the effects of the transmitter filter accuracy, D/A converter, modulator imbalances, untracked phase noise, and PA non-linearity.

Samsung Ex. 1005, Page 460 of 1147

As mentioned earlier, the EVM is a measure of the difference between the reference waveform and the measured waveform. In practice, before calculating the EVM, the measured waveform is corrected by the sample timing offset and RF frequency offset. Then the IQ origin offset is removed from the measured waveform before calculating the EVM. The measured waveform is further modified by selecting the absolute phase and absolute amplitude of the transmit chain. The EVM value is defined after the front-end IDFT as the square root of the ratio of the mean error vector power to the mean reference signal power expressed as a percentage value.

The basic EVM measurement interval in the time-domain depends on the type of physical channel. The measurement interval is equal to the length of one preamble sequence for the physical random access channel (PRACH), and is equal to the length of one slot for the physical uplink control channel (PUCCH) and physical uplink shared channel (PUSCH) in the time-domain. When the PUSCH or PUCCH transmission slot is shortened due to multiplexing with Sounding Reference Signal (SRS), the EVM measurement interval is accordingly reduced by one symbol. The PUSCH or PUCCH EVM measurement interval is also reduced when the mean power, modulation, or allocation between slots is expected to change. In the case of PUSCH transmission, the measurement interval is reduced by a time interval equal to the sum of 5 μs and the applicable exclusion period adjacent to the point where the power change is expected to occur. The PUSCH exclusion period is applied to the signal obtained after the front-end IDFT. In the case of PUCCH transmission with power change, the PUCCH EVM measurement interval is reduced by one symbol adjacent to the point where the power change is expected to occur [1].

9.4.3 Effect of carrier frequency offset on OFDM signals

As mentioned earlier, an OFDM system transmits information as a series of OFDM symbols. The time-domain samples $x_m(n)$ of the mth OFDM symbol are generated by performing IDFT on the information symbols $X_m(k)|_{k=0,1,\ldots,N_{\text{FFT}}-1}$, as follows [132]:

$$x_m(n) = \frac{1}{N_{\text{FFT}}} \sum_{k=0}^{N_{\text{FFT}}-1} X_m(k) e^{j2\pi k(n-N_{\text{CP}})/N_{\text{FFT}}}; \quad \forall 0 \le n \le N_{\text{FFT}} + N_{\text{CP}} - 1 \qquad (9.16)$$

where N_{FFT} and N_{CP} are the number of data samples and cyclic prefix samples, respectively. The OFDM symbol $x_m(n)$ is transmitted through a channel $h_m(n)$ and is perturbed by a Gaussian noise $z_m(n)$. The channel $h_m(n)$ is assumed to be block stationary, i.e., it is time invariant over each OFDM symbol. With this assumption, the output $y_m(n)$ of the channel can be represented as $y_m(n) = h_m(n)^* x_m(n) + z_m(n)$, where $h_m(n)^* x_m(n) = \sum_{k=-\infty}^{+\infty} h_m(k) x_m(k-n)$ and $z_m(n)$ is a zero-mean additive white Gaussian noise (AWGN) with variance σ_z^2.

Since the channel impulse response $h_m(n)$ is assumed to be block stationary, the channel response does not change over each OFDM symbol; however, the channel response $h_m(n)$ may vary across different OFDM symbols, thus it is a function of the OFDM symbol index m.

When the receiver oscillator is not perfectly synchronized to the transmitter oscillator, there can be a CFO $\Delta f_{\text{CFO}} = f_{\text{TX}} - f_{\text{RX}}$ between the transmitter carrier frequency f_{TX} and the receiver carrier frequency f_{RX}. Furthermore, there may be a phase offset θ_0 between the transmitter carrier and the receiver carrier. The mth received symbol $y_m(n)$ can be represented as $y_m(n) = e^{j\{2\pi\Delta f_{\text{CFO}}[n+m(N_{\text{FFT}}+N_{\text{CP}})]T_s+\theta_0\}}[h_m(n)^* x_m(n)] + z_m(n)$, where T_s is the sampling period. The CFO Δf_{CFO} can be represented relative to the sub-carrier bandwidth $(1/N_{\text{FFT}}T_s)$ by defining the relative frequency offset $\delta_{\text{CFO}} = \Delta f_{\text{CFO}} N_{\text{FFT}} T_s$. The CFO attenuates the desired signal and introduces the ICI, thus decreasing the SNR. The SNR of the kth sub-carrier can be expressed as $\text{SNR}_k(\delta_{\text{CFO}}) = (\phi^2_{N_{\text{FFT}}}(\delta_{\text{CFO}})P_h\sigma_x^2/(1 - \phi^2_{N_{\text{FFT}}}(\delta_{\text{CFO}}))P_h\sigma_x^2 + \sigma_z^2)$, where $\phi_{N_{\text{FFT}}}(\delta_{\text{CFO}}) = (\sin(\pi\delta_{\text{CFO}})/N_{\text{FFT}}\sin(\pi\delta_{\text{CFO}}/N_{\text{FFT}}))$ in order to demonstrate the dependence of the SNR on the frequency offset [132]. In the latter equation, P_h, σ_x^2, and σ_z^2 denote the total average power of channel impulse response, variance of the signal, and variance of the additive noise, respectively. The sub-carrier index k is dropped since the SNR is the same for all sub-carriers. From this SNR expression, it is clearly seen that the effect of the frequency offset is to decrease the signal power by $\phi^2_{N_{\text{FFT}}}(\delta_{\text{CFO}})$ and to convert the decreased power to interference power. The SNR depends not only on the frequency offset δ_{CFO}, but also on the number of sub-carriers; however, as N_{FFT} increases $\phi^2_{N_{\text{FFT}}}(\delta_{\text{CFO}})$ converges to $\text{sinc}^2(\delta_{\text{CFO}})$. Therefore, the SNR converges to $\text{SNR}(\delta_{\text{CFO}}) = (\text{sinc}^2(\delta_{\text{CFO}})P_h\sigma_x^2/(1 - \text{sinc}^2(\delta_{\text{CFO}}))P_h\sigma_x^2 + \sigma_z^2)$ as N_{FFT} becomes increasingly large [132]. The effect of CFO on OFDM signal in the frequency-domain, which appears as inter-carrier interference, for different values of δ_{CFO} is shown in Figure 9.17. Figure 9.18 illustrates the variation of $\text{SNR}(\delta_{\text{CFO}})$ as a function of δ_{CFO} with $\gamma = P_h\sigma_x^2/\sigma_z^2$ as a parameter. In the above equations, the power of inter-carrier interference as a function of relative CFO is defined as $P_{\text{ICI}}(\delta_{\text{CFO}}) = (1 - \text{sinc}^2(\delta_{\text{CFO}}))P_h\sigma_x^2$. In practice, the sub-carrier spacing is not the same among different tones due to mismatched oscillators (i.e., frequency offset), Doppler shift, and timing synchronization errors, resulting in inter-carrier interference and loss of orthogonality. It can be seen that the ICI increases with the increase of the OFDM symbol duration (or alternatively decrease of sub-carrier spacing) and the frequency offset. The effects of timing offset are typically less than that of the frequency offset, provided that the cyclic prefix is sufficiently large.

It can be shown that the ICI power can be calculated as a function of the generic Doppler power spectrum $\Lambda(\nu)$ as follows [14]:

$$P_{\text{ICI}} = \int_{-\nu_{\max}}^{\nu_{\max}} \Lambda(\nu)(1 - \text{sinc}^2(T_s\nu))\text{d}\nu \tag{9.17}$$

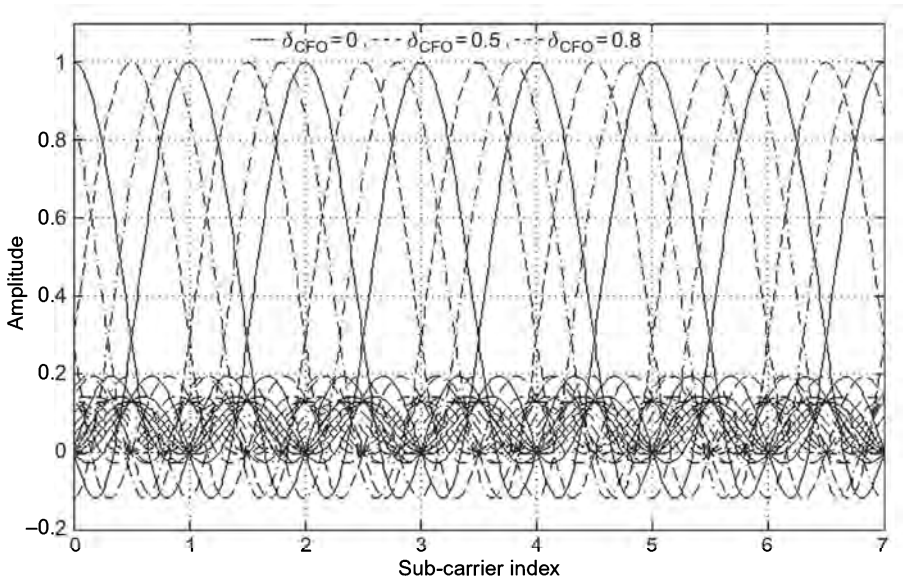


FIGURE 9.17

The effect of carrier frequency offset on OFDM signal for various δ_{CFO} values [159].

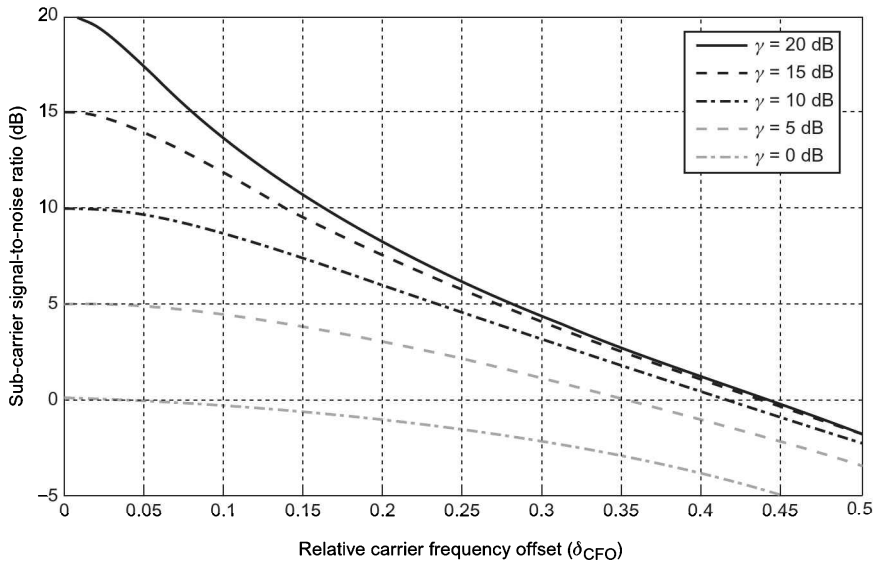


FIGURE 9.18

Variation of the SNR as a function of relative carrier frequency offset δ_{CFO} [159].

where ν_{max} denotes the maximum Doppler frequency. We further assume that the transmitted signal power is normalized. It can be noted that the ICI generated as a result of CFO is a special case of the above equation when $\Lambda(f) = \delta(f - f_{CFO})$, in which $\delta(f)$ represents the Dirac delta function. Using the classic Jakes' model of Doppler spread where the spaced-time correlation function is defined as $\Phi(t) = J_0(2\pi\nu t)$ in which $J_0(x)$ denotes the zero-order Bessel function of the first kind [131], the ICI power can be written as follows:

$$P_{ICI_{Jakes}} = 1 - 2\int_0^1 (1-f)J_0(2\pi\nu_{max}T_s f)\mathrm{d}f \tag{9.18}$$

which approximately gives an upper bound on the ICI power due to Doppler spread. Comparison of the power of ICI generated by CFO and Doppler spread suggests that the ICI impairment due to the former is higher than the latter.

The effect of CFO on EVM for an OFDM system with $E_b/N_0 = 30$ dB, $N_{FFT} = 64$ has also been investigated and is shown in Figure 9.19. It is shown that the EVM increases with increasing values of relative CFO even at such relatively high SNR.

Figure 9.20 shows the bit error rate sensitivity of an OFDM system under AWGN channel to CFO. The increasing values of relative CFO degrade the BER performance of the system.

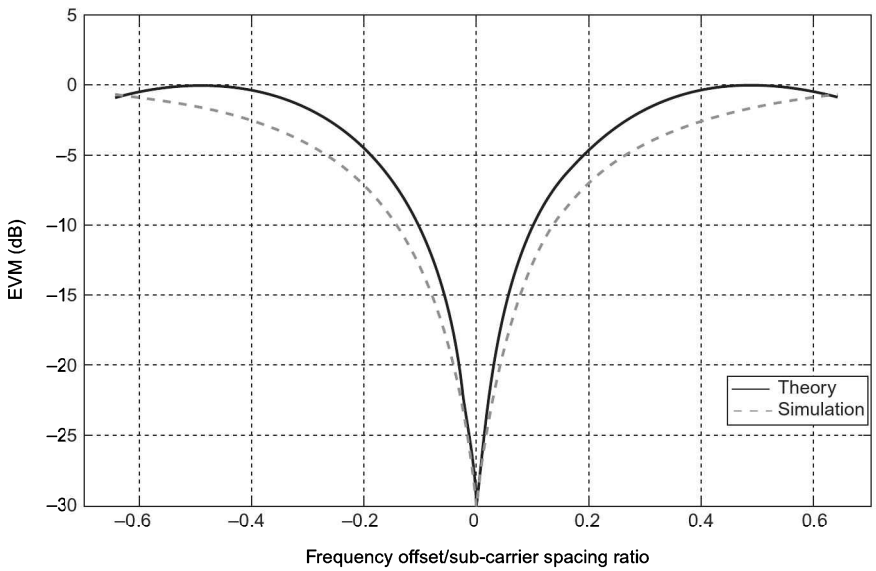


FIGURE 9.19

Error vector magnitude versus frequency offset in OFDM ($E_b/N_0 = 30$ dB, $N_{FFT} = 64$) [159].

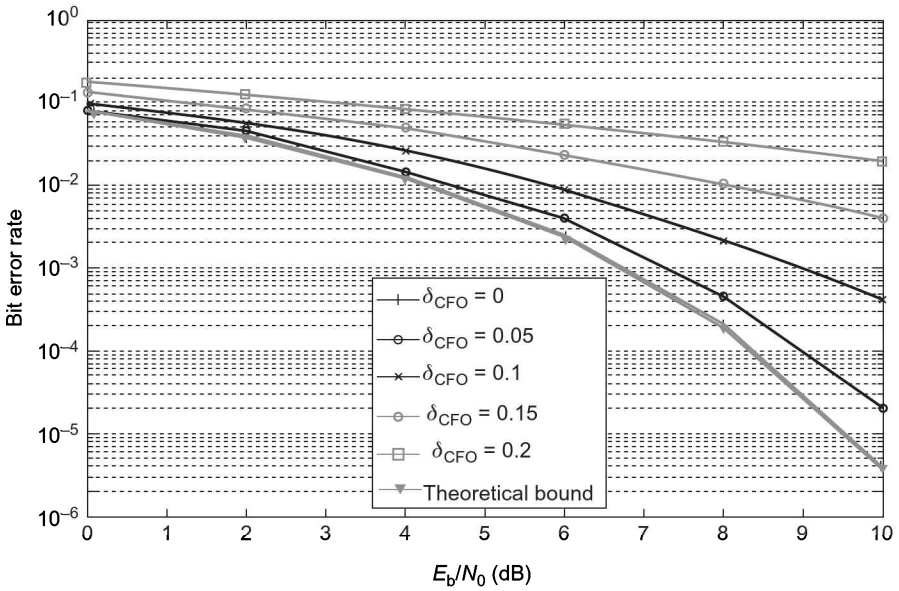


FIGURE 9.20

BER sensitivity to relative CFO δ_{CFO} in an OFDM system in AWGN channel [141].

9.4.4 Effect of oscillator phase noise on OFDM signals

Oscillators are used in typical radio circuits to drive the RF mixers which are utilized for the up-conversion/down-conversion of the band-pass signals. Ideally, the spectrum of the oscillator is expected to have an impulse at the frequency of oscillation with no frequency components elsewhere. However, the spectrum of a practical oscillator's output does exhibit spectrum variation around the oscillation frequency due to phase noise. The impact of local oscillator phase noise on the performance of an OFDM system has been extensively studied in the literature [159]. It has been shown that phase noise may have significant effects on OFDM signals with small sub-carrier spacing (i.e., large OFDM symbol duration in time). Long symbol duration is required for implementing a long guard interval that can mitigate long multipath delay in single frequency operation without excessive reduction of data throughput. The studies suggest that phase noise in OFDM systems can result in two effects: a common sub-carrier phase rotation on all the sub-carriers; and a thermal-noise-like sub-carrier de-orthogonality.

The common phase error, i.e., constellation rotation, on the demodulated sub-carriers is caused by the phase noise spectrum from DC up to the frequency of sub-carrier spacing. This low-pass effect is due to the long integration time of the OFDM symbol duration. This phase error can in principle be corrected by using pilots within the same symbol (in-band pilots). The phase error causes sub-carrier

constellation blurring rather than rotation. It results from the phase noise spectrum contained within the system bandwidth. This part of the phase noise is more crucial, since it cannot be easily corrected. The SNR degradation caused by the common phase error can be quantified as $S/N_{\text{phase-rotation}} = 1/(I(\alpha)\beta\,\Delta f)$, where Δf is the sub-carrier spacing, β denotes the upper bound of the phase noise spectral mask, α is the ratio of the equivalent spectrum mask noise bandwidth and sub-carrier spacing, and $I(\alpha) = \int_{-\alpha}^{+\alpha} \text{sinc}(\pi x)\text{d}x$ with $I(0.5) = 0.774$, $I(1) = 0.903$, and $I(\infty) = 0.774$. It can be seen that, when $\alpha > 1$, the common phase error decreases as the sub-carrier spacing decreases [52].

To mathematically model the effect of the phase noise, let us consider the noisy output of an oscillator which contains phase noise $\phi(t)$ as follows $v(t) = A_c \cos(\omega_c t + \phi(t))$. Let us further assume that the stochastic variation of the phase can be modeled as the output of a system with a step function impulse response and input perturbation $n(t)$ as follows: $\phi(t) = \int_{-\infty}^{t} n(\tau)\text{d}\tau$. Based on the above assumption, the single-sided power spectral density of the phase can be written as follows: $S_\phi(f) = S_n(f)/(2\pi f)^2$, in which $S_n(f)$ denotes the noise power spectral density function. As an example, if $S_n(f)$ is modeled as white noise, then $S_\phi(f) \approx f^{-2}$, and if $S_n(f)$ is modeled as flicker noise, then $S_\phi(f) \approx f^{-3}$. Considering that the power spectral density of the phase is difficult to observe, one may alternatively look at the power spectral density of the oscillator's noisy output $v(t)$. It can be shown that the power spectral density of $v(t)$ can be calculated as follows [159]:

$$S_v(f) = \sum_{k=-\infty}^{k=+\infty} \left(\frac{a_k^2 + b_k^2}{2}\right) \frac{\beta k^2 f_c^2}{\beta^2 \pi^2 k^4 f_c^4 + (f - k f_c)^2} \tag{9.19}$$

where $\{a_k\}$ and $\{b_k\}$ denote the Fourier series coefficients of $v(t)$ and β is a constant. Given that we are only interested in evaluating $S_v(f)$ at f_c, the above equation can be simplified as follows:

$$S_v(f) = \left(\frac{a_1^2 + b_1^2}{2}\right) \frac{\beta f_c^2}{\beta^2 \pi^2 f_c^4 + (f - f_c)^2} \tag{9.20}$$

The above function is a Lorentz distribution.[8] We now define function $\Omega(f)$ as the ratio of noise power in the 1 Hz bandwidth at offset f from center frequency to carrier power which is expressed in dBc/Hz. As theoretically expected, having a higher phase noise in the signal does not increase the total power. A signal with higher phase noise will have smaller power near f_c and will have a broader spectrum around the center frequency. Conversely, a signal with lower phase noise

[8]The Cauchy distribution is a continuous probability distribution which is also known as the Lorentz distribution, Cauchy–Lorentz distribution, or Lorentzian function. It describes the distribution of a random variable that is the ratio of two independent standard normal random variables and has the probability density function defined as $f(x; 0, 1) = 1/[\pi(1 + x^2)]$ [30].

Samsung Ex. 1005, Page 466 of 1147

has a sharper peak at the center frequency with less deviation. Therefore, $\Omega(f)$ can be expressed as follows:

$$\Omega(f) = \frac{1}{\pi}\frac{\gamma}{\gamma^2 + (f - f_c)^2} \quad \gamma = \beta\pi f_c^2 \tag{9.21}$$

It can be shown that $\int_{-\infty}^{+\infty} \Omega(f)\mathrm{d}f = 1$ [159]. Figure 9.21 illustrates the ratio of the noise power to carrier power at $f_c = 500$ Hz for different values of β. It can be seen from the figure, for higher values of β, the spectrum becomes wider with smaller magnitude of the main lobe of the spectrum. Note that a wider main lobe does not increase the total power of the carrier.

In summary, an OFDM system has the following advantages [52]:

- It is sufficiently flexible in meeting various design requirements such as low complexity, bandwidth efficiency and scalability, spectrum shaping, and low sensitivity to various impairments.
- OFDM is well suited to multi-antenna transmission schemes.
- It requires no adaptation to instantaneous channel response for low-order modulations. It is robust to impulse interference and channel variations. Channel estimation is required for high-order QAM modulations.
- As a multi-tone transmission technique, OFDM is less sensitive to shift in sampling time in comparison to serial transmission techniques.

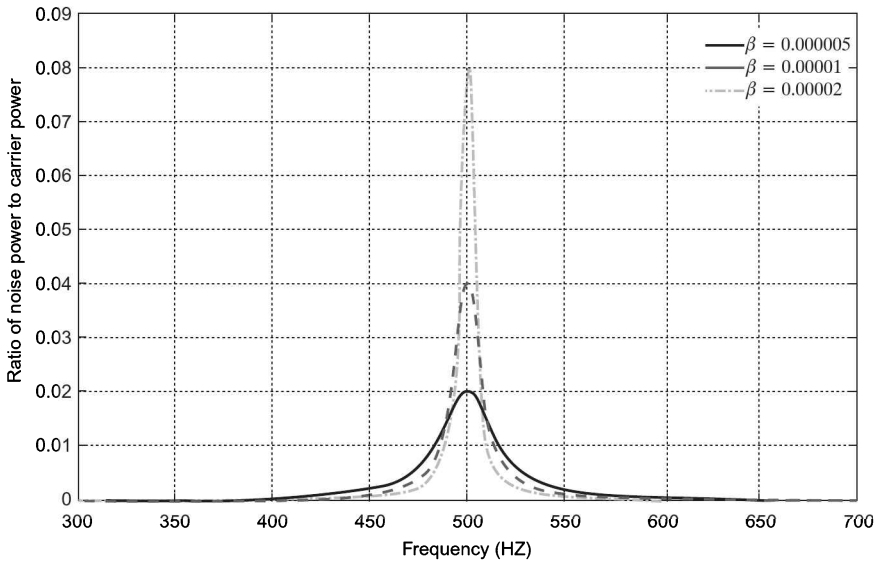


FIGURE 9.21

Plot of the noise power to carrier power $\Omega(f)$ at $f_c = 500$ Hz for different values of β [159].

- The bandwidth efficiency of an OFDM system approaches the Nyquist rate as the FFT size increases (this, however, increases system complexity and is subject to an upper limit due to channel variations and CFO).
- For fixed bandwidth efficiency, the complexity (multiplications per symbol) of an FFT-based OFDM system grows logarithmically with the increase of the channel multipath delay spread, in comparison with a linear increase of complexity for equalizers for a single carrier system.
- A properly coded and interleaved OFDM system can exceed the performance of many other practical systems. Higher transmitter output back-off is required in comparison to a single carrier system, because of the high PAPR of an OFDM system.
- As a parallel transmission technique, OFDM is more sensitive to CFO and tone interference than that of a single carrier system.

9.5 Multiple access schemes

Orthogonal frequency division multiple access (OFDMA) is the multi-user variant of the OFDM scheme where multiple access is achieved by assigning subsets of sub-carriers to different users, allowing simultaneous data transmission from several users. In OFDMA, the radio resources are two-dimensional (2D) regions over time (an integer number of OFDM symbols) and frequency (a number of contiguous or non-contiguous sub-carriers). The difference between OFDM and OFDMA is illustrated in Figure 9.22. Similar to OFDM, OFDMA employs multiple closely spaced sub-carriers, but the sub-carriers are divided into groups of sub-carriers where each group is called a resource block. The grouping of sub-carriers into groups of resource blocks is referred to as sub-channelization. The sub-carriers that form a resource block do not need to be physically adjacent. In the downlink, a resource block may be allocated to different users. In the uplink, a user may be assigned to one or more resource blocks.

Sub-channelization defines sub-channels that can be allocated to mobile stations depending on their channel conditions and service requirements. Using sub-channelization, within the same time slot (i.e., an integer number of OFDM symbols), an OFDMA system can allocate more transmit power to user devices with lower SNR, and less power to user devices with higher SNR. Sub-channelization also enables the base station to allocate higher power to sub-channels assigned to indoor mobile terminals, resulting in better indoor coverage.

There are various multiple access schemes that have been used in cellular systems in the past few years which allow the network to share the available radio resources (i.e., time, frequency, code, and space) among a number of users in the cell in the downlink and uplink directions. Figure 9.23 illustrates the concept of resource sharing in various multiple access schemes. As mentioned earlier, OFDMA has been a promising multiple-access scheme that has recently been

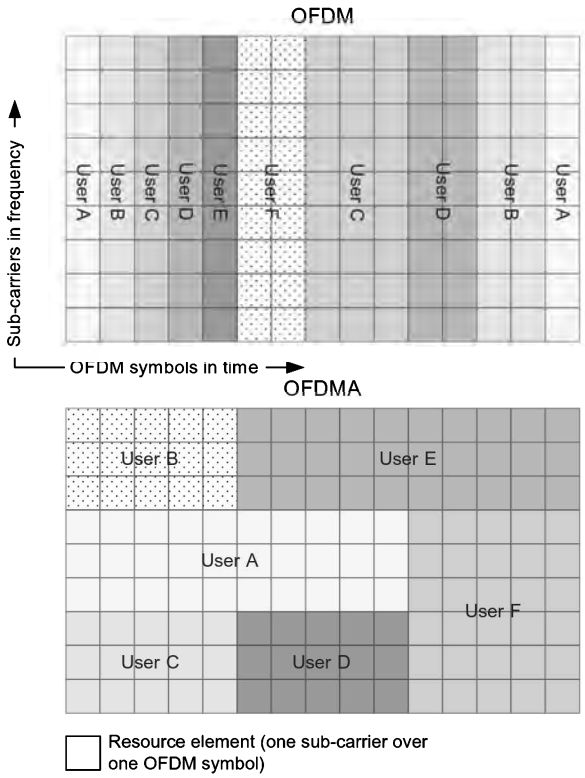


FIGURE 9.22

Illustration of OFDM and OFDMA concepts.

used in mobile broadband radio access technologies. The E-UTRA uses OFDMA and SC-FDMA as the multiple access schemes in the downlink and uplink directions, respectively. The OFDMA/SC-FDMA parameters for the E-UTRA have been chosen to facilitate coexistence with the UTRA on the same platform and sharing the same temporal sampling frequency of 30.72 MHz. The E-UTRA OFDMA/SC-FDMA parameters are given in Table 9.2. The following parameters are compatible with UMTS sampling frequency/chip rate of 3.84 Mcps and frame timing.

The OFDMA/SC-FDMA parameters in E-UTRA satisfy general design criteria for OFDM systems as follows:

$$\begin{aligned} &T_{CP} \geq \tau_{RMS} && \text{(to prevent ISI due to delay spread)} \\ &\Delta f \gg \nu_{max} && \text{(to mitigate ICI due to Doppler spread)} \\ &T_{CP}\ \Delta f \ll 1 && \text{(for spectral efficiency improvement)} \end{aligned} \tag{9.22}$$

Samsung Ex. 1005, Page 469 of 1147

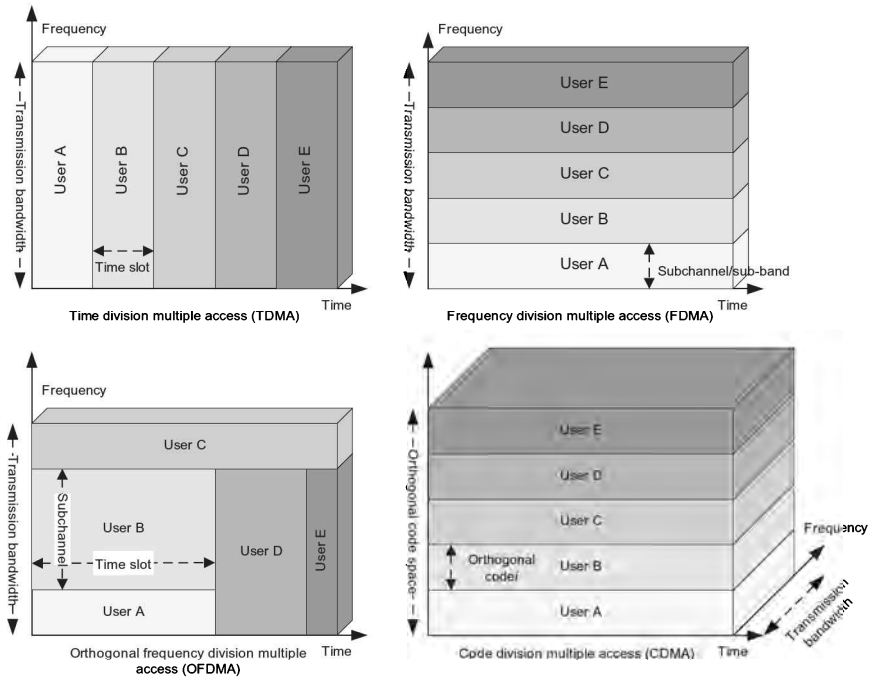


FIGURE 9.23

An illustration of the concept of different multiple access schemes.

In OFDMA, an OFDM symbol is constructed of sub-carriers, the number of which are determined by the FFT size. There are several sub-carrier types: (1) data sub-carriers are used for data transmission; (2) pilot or reference-signal sub-carriers are utilized for channel estimation and coherent detection; and (3) null sub-carriers are not used for pilot/data transmission. The null sub-carriers, including the DC sub-carrier, are used for guard bands. The number of used (or occupied) sub-carriers is always less than the FFT/IFFT size. The guard bands are used to allow spectrum sharing and to reduce the adjacent channel interference and out-of-band emissions. The sampling frequency is selected to be greater than or equal to the channel bandwidth. The number of time samples in a radio frame is always an integer, and to further simplify the design of analog transmit filter the sampling frequency is scaled by a factor greater than one (e.g., the sampling frequency for 20 MHz bandwidth is 30.72 MHz). As given in Table 9.2, the value of the over-sampling factor depends on the channel bandwidth. The E-UTRA supports a large number of channel bandwidths. The channel bandwidths that are integer multiples of 5 MHz are more important. The bandwidths that are not given

Table 9.2 E-UTRA OFDMA/SC-FDMA Parameters [3]

Parameter		Value					
Channel bandwidth (MHz)		1.4	3	5	10	15	20
Number of resource blocks $N_{RB}^{DL/UL}$		6	15	25	50	75	100
Number of occupied sub-carriers		72	180	300	600	900	1200
IFFT/FFT size		128	256	512	1024	1536	2048
Sub-carrier spacing Δf (kHz)		15 (7.5)					
Sampling frequency (MHz)		1.92	3.84	7.68	15.36	23.04	30.72
Samples/slot		960	1920	3840	7680	11520	15360
CP size T_{CP}	Normal CP ($\Delta f = 15$ kHz)	5.21 μs (first symbol of the slot) 4.69 μs (other symbols of the slot) $N_s = 7$ symbols/slot $N_{CP}(l) = \begin{cases} 160 & l = 0 \\ 144 & l = 1, 2, \ldots, 6 \end{cases}$					
	Extended CP ($\Delta f = 15$ kHz)	16.67 μs $N_s = 6$ symbols/slot $N_{CP}(l) = 512 \quad l = 0, 1, \ldots, 5$					
	Extended CP ($\Delta f = 7.5$ kHz)	33.33 μs $N_s = 3$ symbols/slot $N_{CP}(l) = 1024 \quad l = 0, 1, 2$					

in Table 9.2 are supported via tone dropping, i.e., dropping the sub-carriers at the edges of the frequency band based on 10 and 20 MHz systems (e.g., for 6 MHz channel bandwidth, the parameter of 10 MHz is used and the edge sub-carriers are dropped to fit the spectrum block).

9.6 Duplex modes

One of the key elements of any radio communication system is the way in which radio communications are maintained in both directions. For cellular systems, it is necessary to enable transmission of data in both directions simultaneously, and this places a number of constraints on the schemes that may be used to control the transmission flow. As a result, the duplex scheme forms a very basic part of the overall specification for a cellular or any radio communication system. The term duplex refers to the bidirectional communications between two devices. When unpaired spectrum, or alternatively the same RF carrier is used for downlink and uplink communications, the transmit/receive functions are time-multiplexed.

When paired spectrum, or alternatively two RF carriers are used for downlink and uplink communications, the transmit/receive functions are frequency-multiplexed. The FDD is a duplex scheme in which uplink and downlink

transmissions occur simultaneously using different frequencies. The downlink and uplink frequencies are separated by a sufficiently large frequency offset (see Table 9.1 for various E-UTRA bands). For the FDD scheme to properly operate, it is necessary that the frequency separation, i.e., channel separation between the transmission and reception frequencies, be sufficient to prevent the receiver being excessively affected by the transmitter signal. Receiver blocking is an important issue with FDD schemes, and often highly selective filters may be required. For cellular systems using FDD, filters are required within the base station and also the user terminal to ensure sufficient isolation of the transmitter signal without desensitizing the receiver. While implementation cost is not a significant constraint for the base stations, placing a filter into the user terminals is more of an issue. The use of an FDD system does enable simultaneous transmission and reception of signals. However, two channels are required and this may not always use the available spectrum efficiently. The spectrum used for FDD systems is allocated by the regulatory authorities. Since there is a frequency separation between the uplink and downlink frequencies, it is not normally possible to reallocate spectrum to change the balance between the capacity of the uplink and downlink directions, if there are changing capacity requirements for each direction.

The TDD is a duplex scheme where uplink and downlink transmissions occur at different times, but may share the same frequency. In other words, the downlink and uplink transmissions are multiplexed in time and are not concurrent. While FDD transmissions require a large frequency separation between the transmit and receive frequencies, TDD schemes require a guard time or guard interval between transmission and reception. This must be sufficient to allow the signals traveling from the remote transmitter to arrive before a transmission is started and the receiver is shut down. Although this delay is relatively short, when switching between transmission and reception many times a second, even a small guard time can reduce the efficiency of the system, because a percentage of the time must be used for the guard interval. For small-sized cells, e.g., up to a mile or so, the guard interval is normally small and acceptable. For large cell sizes, it may become an issue.

The guard interval required for TDD will comprise two main elements: (1) a time allowance for the propagation delay for any transmission from a remote transmitter to arrive at the receiver. This will depend upon the distances involved (i.e., cell radius). (2) A time allowance for the transceiver to switch from receive to transmit mode. The switching times can vary considerably depending on the implementation, but can take a few microseconds. As a result, TDD is not normally suitable for use over very large cell sizes as the guard time increases and the spectral efficiency decreases.

It is often found that traffic in both directions is not balanced. Typically there is more data transmitted in the downlink direction of a cellular system. This means that, ideally, the capacity should be greater in the downlink direction. Using a TDD system, it is possible to change the capacity in either direction relatively easily by changing the number of time slots allocated to each direction. Often this is dynamically configurable so it can be altered to meet the demand. A further aspect

to be noted with TDD transmission is that of latency. Since the data may not be immediately transmitted, a result of the time multiplexing between transmit and receive circuitry, there may be a small delay between the time that the data is generated until it is actually transmitted in a TDD system. Typically this may be a few milliseconds depending upon the frame timing. Both TDD and FDD have their advantages, and each can be used in different deployment scenarios. Before deciding on a particular type of duplex scheme, it is necessary to analyze the advantages and disadvantages of each duplex mode. Table 9.3 summarizes the relative advantages/disadvantages of TDD and FDD systems.

The E-UTRA supports TDD and FDD schemes with a great extent of commonality in the baseband processing. In order to reduce the implementation complexity and cost of FDD terminals, and further to increase the reuse of baseband functional elements, a half-duplex FDD (H-FDD) operation is supported where the downlink and uplink transmissions are not simultaneous but occur in two different frequencies. As shown in Figure 9.24, classic H-FDD operation does not efficiently utilize the radio resources on the downlink and uplink RF carriers. The complementary grouping and scheduling of users would allow efficient use of downlink and uplink resources in an H-FDD operation. The various duplex schemes are illustrated in Figure 9.24. For H-FDD operation, a guard period (GP) is virtually created by the UE by not receiving the last OFDM symbol(s) of a downlink subframe immediately preceding an uplink subframe in which the UE is active. The length of the GP is the sum of the maximum round-trip propagation delay in the cell, transmit-to-receive, and receive-to-transmit switching delay at the UE.

In E-UTRA, the differences between TDD and FDD schemes are mainly in the physical layer, and therefore are transparent to the higher layers. As a result, there are no significant operational differences between the two modes in the system architecture. At the physical layer, the fundamental design goal was to achieve as much commonality between the two modes as possible. The key design differences between the two stem from the need to support various TDD UL/DL allocations and provide coexistence with other TDD systems. In this regard, several additional features are exclusively developed for E-UTRA TDD. For instance, the FDD and TDD modes differ in the time placement of the downlink synchronization signals. In FDD, the primary and secondary synchronization signals are contiguously placed within one subframe, while in TDD, the two signals are placed in different subframes, separated by two OFDM symbols. However, this separation of the primary and secondary synchronization signals does not affect the performance of synchronization channels relative to FDD systems.

The system-level simulations also show that the sector throughput and cell-edge throughput of E-UTRA TDD and FDD mode are quite similar for best effort traffic when comparing a 10 + 10 MHz FDD system to a 20 MHz TDD system. The spectral efficiency of the TDD mode is slightly degraded due to the

Table 9.3 Comparison of TDD and FDD Attributes

Attribute	TDD	FDD
Paired spectrum	Does not require paired spectrum as both transmit and receive occur on the same channel.	Requires paired spectrum with sufficient frequency separation to allow simultaneous transmission and reception.
Hardware cost	Lower cost as no duplexer[a] is needed to isolate the transmitter and receiver.	Duplexer is needed and the implementation cost is higher
Channel reciprocity	Channel propagation is the same in both directions which enables estimation of the downlink channel from the uplink channel.	Channel characteristics are different in both directions due to use of different frequencies.
UL/DL asymmetry	It is possible to dynamically change the UL and DL ratio based on the traffic volume in each direction.	UL/DL bandwidths are determined by frequency allocation designated by the regulatory authorities. It is therefore not possible to dynamically adapt to the traffic volume.
Guard period/ frequency separation	Guard period required to ensure uplink and downlink transmissions do not interfere. Large guard period will limit the capacity. Larger guard period is normally required if distances are increased to accommodate larger propagation delays. Note that a guard band in frequency-domain is required to suppress interference to adjacent bands.	Frequency separation is required to provide sufficient isolation between uplink and downlink. However, large frequency separation does not impact the capacity. Note that a guard band in frequency-domain is required to suppress interference to adjacent bands.
Discontinuous transmission	Discontinuous transmission is required to allow both uplink and downlink transmissions. This can degrade the performance of the RF power amplifier in the transmitter.	Continuous transmission is possible.
Switching point synchronization	Base stations are required to be synchronized with respect to the uplink and downlink transmission times. If neighboring base stations use different uplink and downlink assignments and share the same channel, then interference may occur between cells.	Not applicable.

[a] *A diplexer is a passive device that implements frequency domain multiplexing. Two ports are multiplexed onto a third port. The signals on each input port occupy non-overlapping frequency bands. Consequently, the signals on the input ports can coexist on the output port without interfering with each other. On the other hand, a duplexer is a device that allows bidirectional communication over a single channel. In radar and radio communications systems, the duplexer isolates the receiver from the transmitter while permitting them to share a common antenna. Most radio repeater systems include a duplexer. Duplexers are designed for operation in the frequency band used by the receiver and transmitter, and must be capable of handling the output power of the transmitter. They must provide sufficient isolation between transmitter and receiver to prevent receiver desensitization [34].*

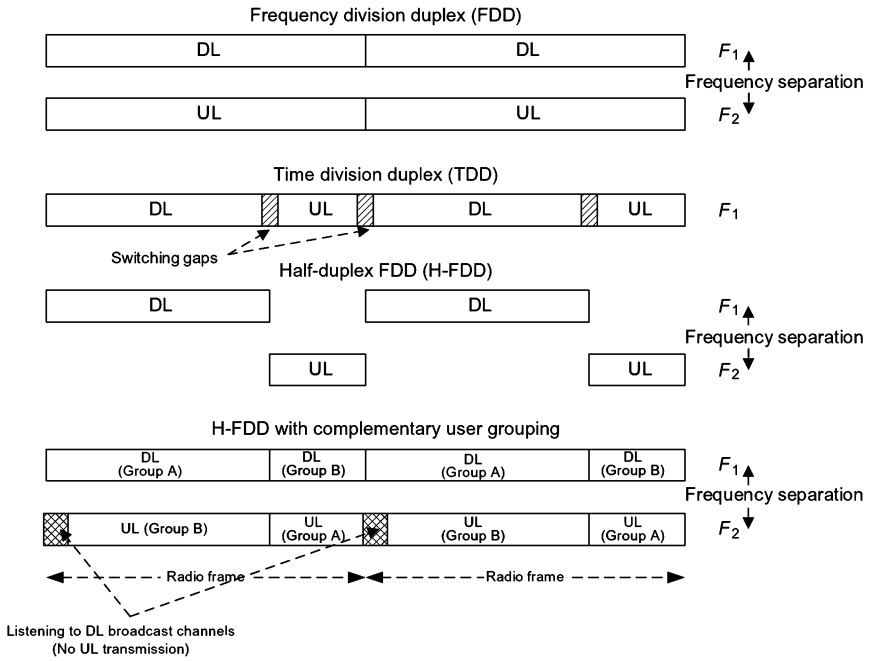


FIGURE 9.24

Illustration of the concept of different duplexing schemes.

additional overhead corresponding to the GP. Another advantage of TDD systems is the flexibility they may offer to network operators in terms of adjusting the DL/UL ratio. This feature allows operators to configure the DL/UL ratio to suit the user traffic in their network.

A common misperception regarding TDD systems is that the uplink coverage of the TDD systems is inferior relative to their FDD counterpart. The link-budget analysis for both systems in the outdoor and indoor environments with compatible system configuration and transmitter and receiver settings shows that the TDD system matches the uplink coverage of the FDD counterpart [10]. In fact, the E-UTRA TDD and FDD modes utilize the same slot/subframe structure and resource block size, thus the same amount of power is transmitted during the same time interval, implying that the coverage will be the same. What has been fueling this misconception is the single-user uplink data rate in an unloaded network at the cell-edge, where an FDD system will show a cell-edge performance that is twice that of the TDD system. It must be noted that the link budget analysis focuses only on the cell-edge single user performance when that user terminal is at the maximum limit of its uplink power, hence it can only be given access to five resource blocks (one resource block every 1 ms for 5 ms, or half of the length of

Samsung Ex. 1005, Page 475 of 1147

the 10 ms E-UTRA radio frame with DL:UL = 1:1 ratio) to meet the 200 mW maximum power requirement. In comparison, in an FDD system at the cell-edge, one single device will be able to send 200 mW continuously for 10 ms, assuming that the scheduler can give the device access to 10 resource blocks of 1 ms with no waiting time. However, in real-life scenarios, the cells have more than one user and the performance of FDD and TDD systems will be very similar, as the FDD device will not be able to access the full-frame 10 uplink resource blocks similar to an unloaded network. In that case, because the TDD channel is twice as wide as the FDD channel, the user is likely to be given access to as many resource blocks as an FDD system despite the half-uplink time in the frame compared to an FDD frame. For network operators, an E-UTRA TDD will provide the same coverage and very similar sector throughput relative to an E-UTRA FDD system. The TDD with its adaptable DL:UL ratio would allow an optimized and efficient use of the uplink and downlink bandwidths as the operators are able to match the asymmetry of data traffic and to maximize spectrum utilization and efficiency in almost all scenarios [163].

9.7 Frame structure

Downlink and uplink transmissions are structured into radio frames with 10 ms duration. The E-UTRA supports two frame structures, namely, type 1 and type 2 for FDD and TDD duplex schemes, respectively. The radio frames are numbered using a system frame number (SFN). The SFN is defined using a string of 10 bits which provide an integer value from 0 to 1023. Therefore, the SFN repeats once every 10.24 s, i.e., 1024 frames. The eight most significant bits of the SFN are broadcast within the master information block (MIB) of the system information on the broadcast channel (BCH). The remaining bits of SFN are deduced from the four frame cycles used to transmit the MIB (i.e., the first radio frame corresponds to "00" and the fourth radio frame corresponds to "11"). As given in Table 9.2, using the normal cyclic prefix, there are seven OFDM symbols within each slot. In the case of multi-cell aggregation, the same frame structure is used across all serving cells. Transmissions in multiple cells can be aggregated up to four secondary cells (in this context cell refers to RF carriers) in addition to the primary cell.

To provide enhanced multicast and broadcast services, E-UTRA has the capability to transmit multicast and broadcast content over a single frequency network (MBSFN), where a time-synchronized common waveform is transmitted from multiple eNBs, allowing macro-diversity combining of multi-cell transmissions at the UE. The cyclic prefix is utilized to cover the difference in the propagation delays, which makes the MBSFN transmission appear to the UE as a transmission from a single large cell. In LTE, transmission on a dedicated carrier for MBSFN

with the possibility to use a longer CP with a sub-carrier bandwidth of 7.5 kHz is supported, as well as transmission of MBSFN on a carrier with both MBMS and unicast transmissions using time division multiplexing [7].

9.7.1 Frame structure type 1

Frame structure type 1 is applicable to both full-duplex FDD and H-FDD schemes. In frame structure type 1, as illustrated in Figure 9.25, each 10 ms radio frame is divided into 10 equally-sized subframes ($T_{\text{frame}} = 307,200T_s = 10T_{\text{subframe}}$). Each subframe consists of two equally sized and consecutive 0.5 ms slots ($T_{\text{slot}} = 15360T_s$). Each slot in turn consists of a number of OFDM symbols which can be either seven (normal cyclic prefix) or six (extended cyclic prefix). As shown in Figure 9.25, the useful symbol time is $T_u = 2048T_s = 66.7\ \mu s$. For a normal cyclic prefix, the first symbol has a cyclic prefix lengths of $T_{CP} = 160T_s = 5.2\ \mu s$. The remaining six OFDM symbols have a cyclic prefix of length $T_{CP} = 144T_s = 4.7\ \mu s$. The reason for different cyclic prefix lengths of the first symbol is to make the overall slot length in terms of time units divisible by 15,360. For the extended mode, the cyclic prefix is $T_{CP} = 512T_s = 16.67\ \mu s$. The length of the cyclic prefix is longer than the delay spread of a few microseconds typically encountered in practice. The normal cyclic prefix is used in urban cells and high data rate applications, while the extended cyclic prefix is used in special cases like multi-cell broadcast and in very large cells (e.g., rural areas, low data rate

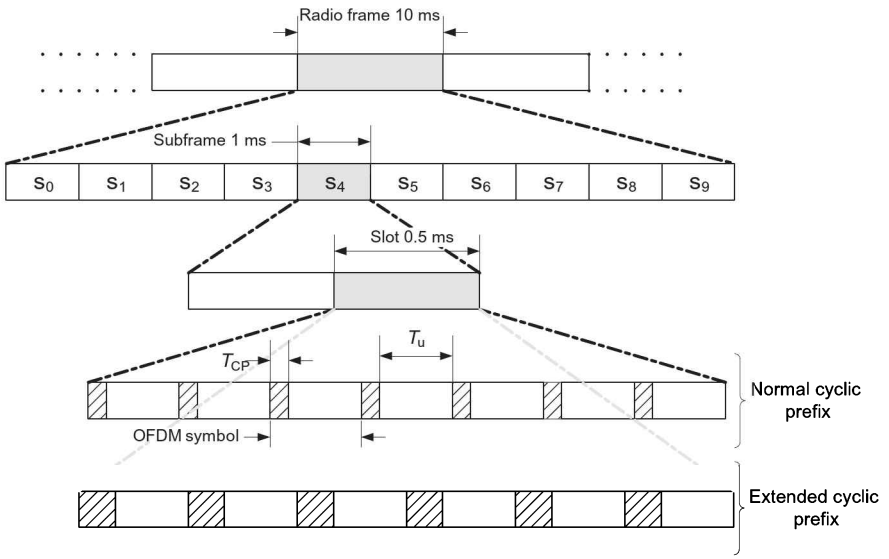


FIGURE 9.25

Frame structure type 1 with different cyclic prefix sizes [3,10].

applications). The cyclic prefix reduces the physical layer capacity by approximately 7.5% in the case of a normal cyclic prefix.

One way to reduce the relative overhead due to cyclic prefix insertion is to reduce the sub-carrier spacing with a corresponding increase in the OFDM symbol duration. However, this will increase the sensitivity of the OFDM transmission to frequency deviations resulting from fast channel variations and large Doppler spread, as well as different types of frequency errors due to RF circuitry. For FDD systems, 10 subframes are separately available for downlink and uplink transmissions in each 10 ms interval. Uplink and downlink transmissions are separated in the frequency-domain. In H-FDD operation, the UE cannot transmit and receive at the same time, while there are no such restrictions in full-duplex FDD.

The transmission time interval (TTI) of the E-UTRA downlink/uplink is 1 ms. The TTI is defined as the duration of the transmission of the physical layer encoded packets over the radio air-interface. In other words, the TTI refers to the length of an independently decodable transmission on the air-link. The data on a transport channel is organized into transport blocks. In each TTI, at most one transport block of variable size is transmitted over the radio air-interface to or from a terminal in the absence of spatial multiplexing. In the case of spatial multiplexing, up to two transport blocks per TTI are transmitted [3].

In the frequency-domain, the number of sub-carriers N_{FFT} ranges from 128 to 2048, depending on the channel bandwidth, with 512 and 1024 for 5 and 10 MHz, respectively, being most commonly used in practice. The sub-carrier spacing is 15 kHz. The sampling rate is chosen as $f_s = 15,000N_{\text{FFT}}$, resulting in a sampling rate that is an integer multiple or sub-multiple of the WCDMA chip rate of 3.84 Mcps. The OFDM parameters have been chosen such that FFT sizes and sampling rates are easily scalable for all operation bandwidths, while at the same time facilitating the implementation of multi-mode terminals with a common reference clock. Not all the sub-carriers are modulated or used. The DC sub-carrier is not used, as well as a number of sub-carriers on either side of the frequency band; thereby, approximately 10% of sub-carriers are used as guard sub-carriers.

9.7.2 Frame structure type 2

Frame structure type 2 is applied to a TDD duplex scheme where the downlink and uplink transmissions are time-multiplexed over the extent of a radio frame (Figure 9.26). Each radio frame of length 10 ms ($T_{\text{frame}} = 307,200T_s$) comprises two half-frames, where each half-frame consists of five subframes of length 1 ms. The supported downlink/uplink ratios (i.e., the ratio of downlink and uplink subframes in a frame) are listed in Table 9.4 and shown in Figure 9.27, where for each subframe in a radio frame, "DL" denotes the subframe reserved for downlink transmissions, "UL" denotes the subframe designated to uplink transmissions, and "S" denotes a special subframe comprising three fields downlink pilot time slot (DwPTS), GP, and uplink pilot time slot (UpPTS). The length of DwPTS and

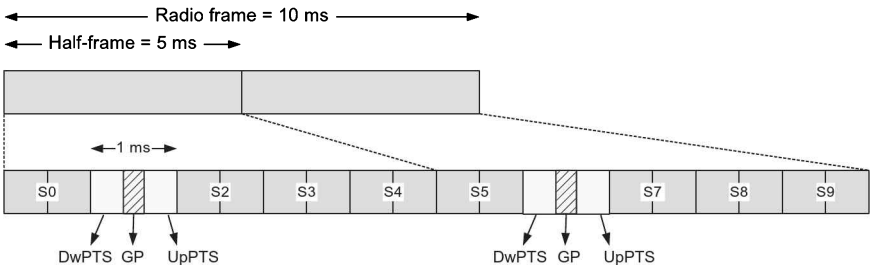


FIGURE 9.26

Frame structure type 2 [3].

Table 9.4 E-UTRA TDD Permissible DL/UL Ratios [3]

E-UTRA TDD Frame Configuration	Switching-Point Periodicity (ms)	Subframe Number									
		0	1	2	3	4	5	6	7	8	9
0	5	DL	S	UL	UL	UL	DL	S	UL	UL	UL
1	5	DL	S	UL	UL	DL	DL	S	UL	UL	DL
2	5	DL	S	UL	DL	DL	DL	S	UL	DL	DL
3	10	DL	S	UL	UL	UL	DL	DL	DL	DL	DL
4	10	DL	S	UL	UL	DL	DL	DL	DL	DL	DL
5	10	DL	S	UL	DL	DL	DL	DL	DL	DL	DL
6	5	DL	S	UL	UL	UL	DL	S	UL	UL	DL

UpPTS are given in Table 9.5, assuming that the total length of DwPTS, GP, and UpPTS is always 1 ms. Each subframe i is defined as two slots $2i$ and $2i + 1$ of length 0.5 ms.

The E-UTRA TDD supports downlink/uplink ratios with both 5 and 10 ms downlink/uplink switching-point periodicity. In the case of 5 ms switching-point periodicity, there are two special subframes within one radio frame. Subframes 0 and 5 and DwPTS are always reserved for downlink transmission due to essential downlink physical channel mapping. The UpPTS interval and the subframe immediately following the special subframe are always reserved for uplink transmission. In case multiple cells are aggregated, and in order to minimize inter-cell interference, the UE may assume the same downlink/uplink ratio across all the cells and that the GPs of the special subframes in different cells have a minimum overlap of $1456T_s$. For the switching from uplink to downlink transmission, no special subframe is provisioned, but the GP includes the sum of switching times from DL to UL and UL to DL. The switching from UL to DL is achieved by provisioning an appropriate timing advance at the UE.

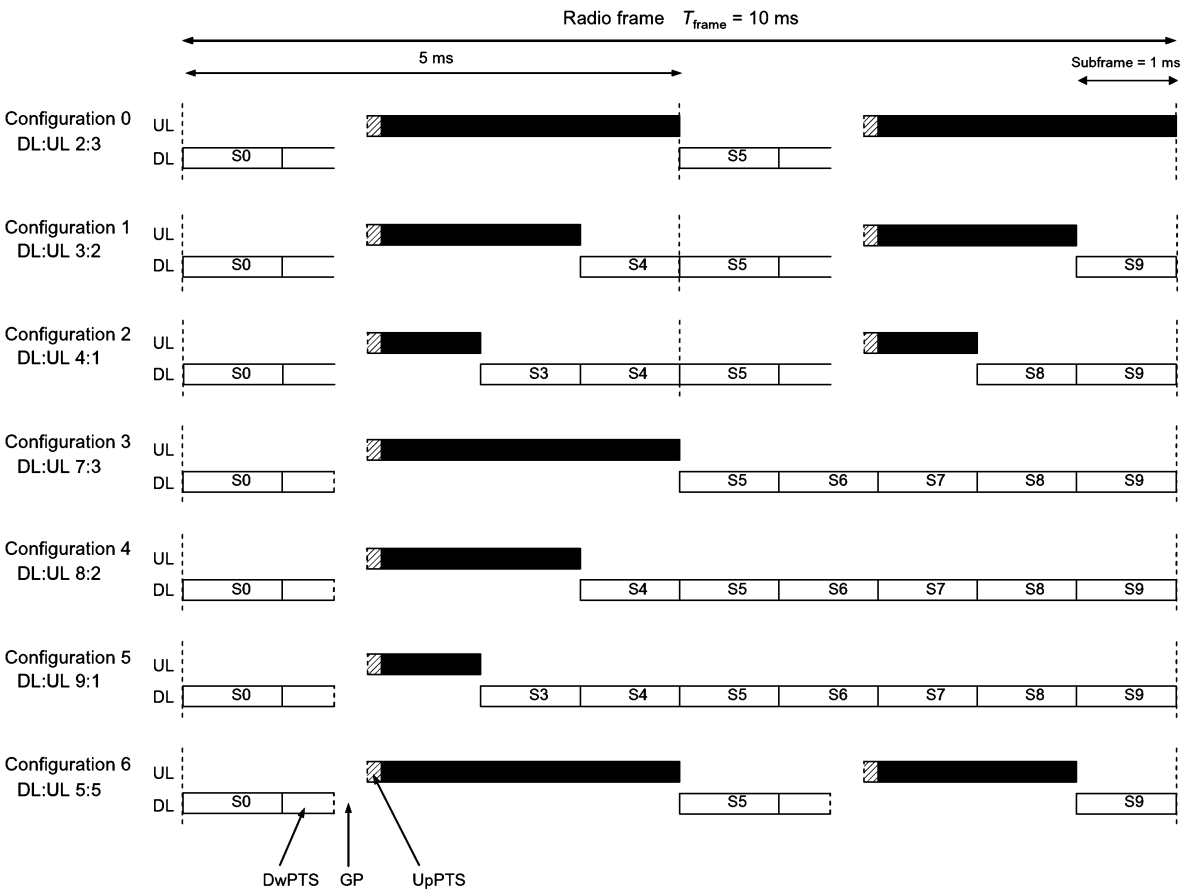


FIGURE 9.27

Illustration of various frame structure type-2 configurations [10].

Table 9.5 Various Special Subframe Configurations (DwPTS/GP/UpPTS) [3]

Special Subframe Configuration	Normal Cyclic Prefix in Downlink				Extended Cyclic Prefix in Downlink			
	DwPTS (Duration/ Number of OFDM Symbols)	GP (Number of OFDM Symbols)	UpPTS		DwPTS (Duration/ Number of OFDM Symbols)	GP (Number of OFDM Symbols)	UpPTS	
			Normal Cyclic Prefix in Uplink (Duration/ Number of OFDM Symbols)	Extended Cyclic Prefix in Uplink (Duration/ Number of OFDM Symbols)			Normal Cyclic Prefix in Uplink (Duration/ Number of OFDM Symbols)	Extended Cyclic Prefix in Uplink (Duration/ Number of OFDM Symbols)
0	$6592T_s$ 3	10			$7680T_s$ 3	8		
1	$19760T_s$ 9	4			$20480T_s$ 8	3		
2	$21952T_s$ 10	3	$2192T_s$ 1	$2560T_s$ 1	$23040T_s$ 9	2	$2192T_s$ 1	$2560T_s$ 1
3	$24144T_s$ 11	2			$25600T_s$ 10	1		
4	$26336T_s$ 12	1			$7680T_s$ 3	7		
5	$6592T_s$ 3	9			$20480T_s$ 8	2		
6	$19760T_s$ 9	3			$23040T_s$ 9	1	$4384T_s$ 2	$5120T_s$ 2
7	$21952T_s$ 10	2	$4384T_s$ 2	$5120T_s$ 2	$12800T_s$			
8	$24144T_s$ 11	1			– –	–	– –	– –
9	$13168T_s$ 6	6			– –	–	– –	– –

Within a radio frame, an E-UTRA TDD system switches multiple times between downlink and uplink transmissions. In this process, different signal propagation times between eNB and various UEs must be taken into consideration in order to prevent interference with other subframes. The timing advance feature prevents conflicts as a result of switching from uplink to downlink transmission. The UEs are informed by the eNB as to when they must start their uplink transmission. The greater the distance between the eNB and the UE, the earlier the UE starts transmitting. This helps to ensure that all signals reach the eNB with proper timing. When switching from the downlink to the uplink, a GP is inserted between the DwPTS and the UpPTS fields. The duration of the GP depends on the signal propagation round-trip time from the eNB to the UE, as well as on the time the UE requires to switch from receiving to transmitting states. The duration of the GP is configured by the network based on the cell size. Because the overall length of the special subframe remains constant, and the GP length varies based on cell size, the lengths of the DwPTS and UpPTS have to be adjusted. There are ten different special subframe configurations specified for E-UTRA TDD, as given in Table 9.5. The length of the GP determines the maximum supportable cell size. The E-UTRA should support cell sizes of up to 100 km, requiring a GP of approximately 666.7 μs. This is possible by choosing configuration 0 in Table 9.5 for the special subframe.

While the GP separates uplink and downlink transmissions, the other special fields are used for data/control signaling transmission. The DwPTS field carries synchronization and user data, as well as the downlink control channel for transmitting scheduling and control information. The UpPTS field is used for transmitting PRACH and SRS.

An example of coexistence with a legacy TD-SCDMA system [171] is shown in Figure 9.28, which illustrates switching point alignment between the two systems. It must be noted that in order to reduce the number of special subframe patterns in LTE TDD, only few legacy TD-SCDMA frame configurations are supported; therefore, the support is limited to the most common deployment scenarios, e.g., TD-SCDMA DL:UL = 5:2 and DL:UL = 4:3 configurations [163]. Despite the unequal size of the slots in TD-SCDMA and LTE TDD, the use of an

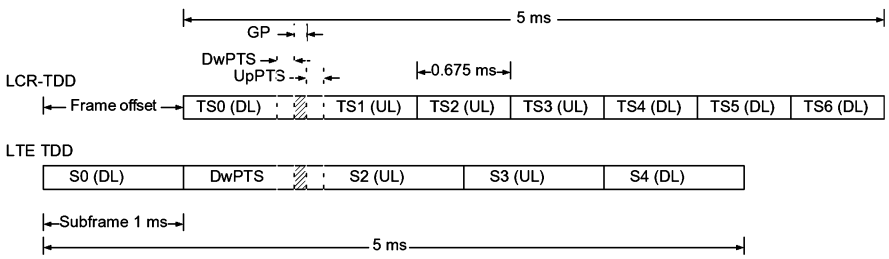


FIGURE 9.28

Example of LTE TDD coexistence with a TD-SCDMA system [163].

appropriate frame offset (as shown in Figure 9.28) allows alignment of the special subframes to the possible extent such that uplink and downlink transmissions of the two systems are synchronized [163].

9.7.3 MBSFN subframe structure

In E-UTRA frame structure, each downlink subframe and the DwPTS fields in the case of TDD is divided into a control region, consisting of the first OFDM symbols of the subframe, and a data region, comprising the remaining OFDM symbols of the subframe. A subset of the downlink subframes in a radio frame on a carrier supporting physical downlink shared channel (PDSCH) transmission can be configured as MBSFN subframes by higher layers. Each MBSFN subframe is divided into a non-MBSFN region and an MBSFN region. The non-MBSFN region spans the first one or two OFDM symbols in an MBSFN subframe. The MBSFN region in an MBSFN subframe is defined as the OFDM symbols not used for the non-MBSFN region.

The MBSFN subframes were originally designed to support MBSFN transmissions. However, they have also been found useful in other areas such as E-UTRA relaying functionality or the new carrier types. The MBSFN subframes are viewed as a generic tool, not necessarily related to MBSFN transmissions, which can provide time–frequency resources for enabling non-backward compatible functionalities relative to LTE Rel-8. An MBSFN subframe, as illustrated in Figure 9.29, consists of a control region of length one or two OFDM symbols, which is identical to normal subframes, followed by an MBSFN data region whose content will depend on the usage of the MBSFN subframe. The reason for maintaining the same control region in the MBSFN subframes is in consideration for normal

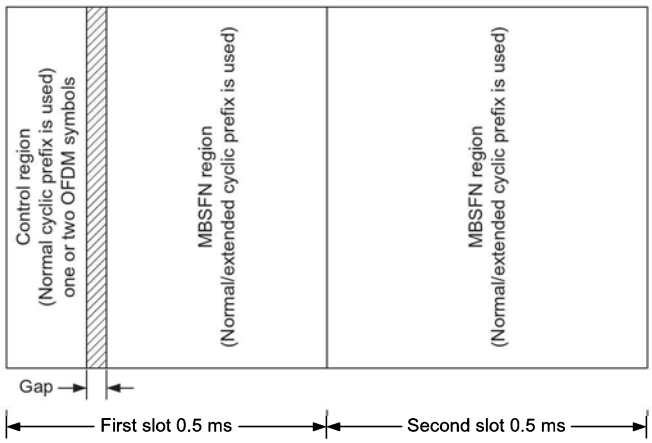


FIGURE 9.29

MBSFN subframe structure [15].

operation of LTE Rel-8 UEs, and in order to transmit control signaling necessary for uplink transmissions. All user terminals will be capable of receiving the control region of an MBSFN subframe. This is the main reason for which MBSFN subframes have been found useful for introducing new and non-backward compatible types of signaling and transmission modes in the new releases of LTE, without compromising the backward compatibility with earlier releases. The user terminals that are not capable of receiving transmissions in the MBSFN data region will simply ignore those transmissions.

Information about the set of subframes that are configured as MBSFN subframes in a cell is provided as part of the system information. In principle, an arbitrary pattern of MBSFN subframes can be configured with the pattern repeating after 40 ms. However, since physical signals and other essential information for system operation need to be transmitted in an orderly fashion to the UEs to find and attach to a cell, subframes where such information is mapped cannot be configured as MBSFN subframes. Therefore, subframes 0, 4, 5, and 9 for FDD and subframes 0, 1, 5, and 6 for TDD cannot be configured as MBSFN subframes, leaving the remaining six subframes as candidates for MBSFN subframes.

9.8 Resource structure, allocation, mapping

The minimum time–frequency resource unit used for downlink transmission is a resource element, which is defined as one sub-carrier over one OFDM symbol. For both TDD and FDD duplex schemes, a group of 12 sub-carriers contiguous in frequency over one slot in time form a PRB, i.e., a two-dimensional region corresponding to one slot in the time-domain and 180 kHz (i.e.,12×15 kHz) in the frequency-domain. The transmissions are allocated in units of PRB. More specifically, the data or control signaling in each slot is transmitted via one or several resource elements of $12 \times N_{RB}^{DL}$ sub-carriers over N_s^{DL} OFDM symbols. The value of the parameter N_{RB}^{DL} depends on the downlink transmission bandwidth configured for the cell, and is given as $\min(N_{RB}^{DL}) \leq N_{RB}^{DL} \leq \max(N_{RB}^{DL})$, where $\min(N_{RB}^{DL}) = 6$ for 1.4 MHz RF channels, and $\max(N_{RB}^{DL}) = 110$ for 20 MHz channels. The set of permissible values for N_{RB}^{DL} are given in Table 9.2. The number of OFDM symbols in a slot N_s^{DL} depends on the cyclic prefix length and the sub-carrier spacing configuration as given in Table 9.2.

There are a number of physical channels and physical signals that are mapped to predetermined (or configurable) locations in time and frequency over a radio frame that are used to assist proper operation of the system and the mobile terminals. The downlink physical channels correspond to a set of resource elements carrying data/control signaling information originating from higher protocol layers. The following downlink physical channels are defined in LTE/LTE-Advanced:

- Physical broadcast channel (PBCH): It carries the MIB portion of the system information. It is mapped to six resource blocks which are centered around

DC sub-carrier in subframe 0 and utilizes the resource elements which are not reserved for transmission of reference signals or other downlink control signaling. The BCH transport block is mapped to four subframes within a 40 ms interval.

- Physical control format indicator channel (PCFICH): It carries the size of PDCCH and is mapped to the first OFDM symbol in each downlink or special subframe. It contains the information on the number of OFDM symbols used for the downlink control region. The exact location of PCFICH is determined by the cell identity and system bandwidth.
- Physical hybrid-ARQ indicator channel (PHICH): It carries HARQ feedback corresponding to uplink data transmissions.
- Physical downlink control channel (PDCCH): It informs the UE(s) about transport format, resource allocation, HARQ feedback, and uplink grants. It is mapped to the first $L = 1, 2, \text{or } 3$ OFDM symbols in each downlink subframe. The number of the OFDM symbols for PDCCH is obtained via PCFICH. Multiple PDCCHs may be monitored by a UE. It also carries uplink power control commands.
- Physical downlink shared channel (PDSCH): It carriers downlink data transmission for different users as well as upper layer signaling.
- Physical multicast channel (PMCH): It is used to broadcast MBMS services.

In addition to the above physical channels, a set of downlink physical signals corresponding to a predetermined set of resource elements are specified and used by the physical layer which do not carry information originating from higher layers but are extremely important for proper LTE/LTE-Advanced system operation. The downlink physical signals include a variety of reference signals (also known as pilots which allow the UE to estimate the communication channel and coherently detect/decode data and control information that can be either cell-specific or UE-specific) and synchronization signals that provide timing and frequency synchronization, as well as physical identity of the cell to UEs attempting to select/reselect the cell. Figure 9.30 depicts the location of the downlink physical channels and signals in time and frequency within a type 1 frame structure.

9.8.1 Physical and virtual resource blocks

Each resource element is uniquely identified by the index pair (k, l) in a slot where $k = 0, \ldots, 12 \times N_{\text{RB}}^{\text{DL}} - 1$ and $l = 0, \ldots, N_{\text{s}}^{\text{DL}} - 1$ are the indices corresponding to frequency and time dimensions, respectively. Throughout this book, we assume that LTE/LTE-Advanced sub-carrier spacing is 15 kHz and there are 12 sub-carriers in each physical and virtual resource block (VRB) per slot. The structure of resource blocks in the downlink is shown in Figure 9.31. LTE/LTE-Advanced defines two types of resource blocks: physical and VRBs where in the latter, the resource blocks over a slot are permuted across frequency dimension to

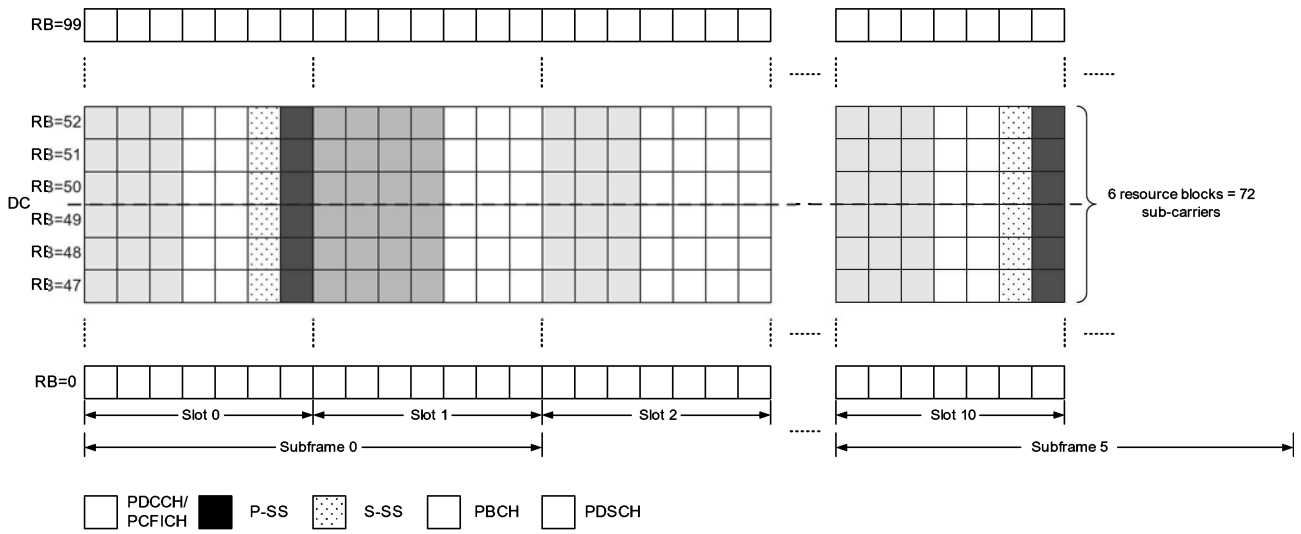


FIGURE 9.30

Physical resource structure and channel mapping (FDD, $N_s^{DL} = 7$, $N_{RB}^{DL} = 100$) [158].

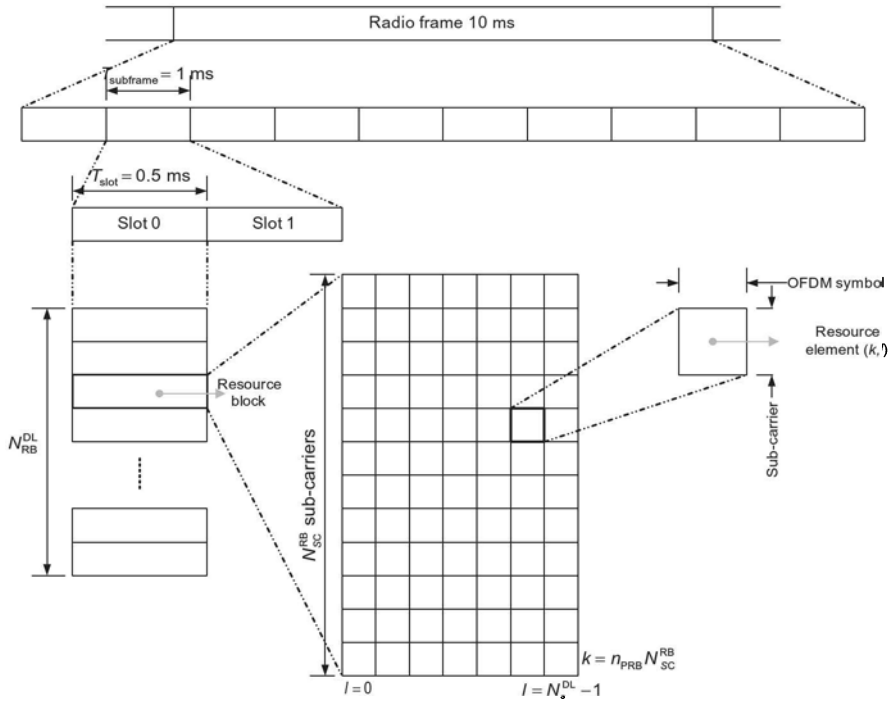


FIGURE 9.31

Resource block structure in time and frequency dimensions [3,162].

take advantage of frequency diversity. Another unit of resource allocation in the downlink is resource block group (RBG), which is a group of resource blocks. Allocating resources in RBG quanta reduces signaling overhead by reducing the number of bits required to uniquely address and identify each RBG relative to individual resource block identification/addressing. The size and number of RBGs depend on the channel bandwidth.

A PRB is defined as N_s^{DL} consecutive OFDM symbols in the time-domain and 12 (or 24 in the case of 7.5 kHz sub-carrier spacing) consecutive sub-carriers in the frequency-domain. A PRB consists of $12 \times N_s^{DL}$ resource elements corresponding to one slot in the time-domain and 180 kHz in the frequency-domain. The PRBs are numbered from 0 to $N_{RB}^{DL} - 1$ in the frequency-domain (see Figure 9.31). The relation between the PRB number n_{PRB} in the frequency-domain and resource elements (k, l) in a slot is given by $n_{PRB} = \lfloor k/12 \rfloor$.

A VRB has the same size as a PRB. There are two types of VRBs: VRBs of localized type and VRBs of distributed type. For each type of VRB, a pair of VRBs over two slots in a subframe is assigned together by a single VRB number n_{VRB}. VRBs of localized type are mapped directly to PRBs such that VRB index

n_{VRB} is identical to PRB index n_{PRB}. The VRBs are numbered from 0 to $N_{VRB}^{DL} - 1$, where $N_{VRB}^{DL} = N_{RB}^{DL}$. In the localized allocations, there is a one-to-one mapping between virtual and PRBs, whereas in the distributed case of resource allocation, the VRB numbers are mapped to PRB numbers according to a certain rule and inter-slot hopping is further utilized. In the latter, the first part of a VRB pair is mapped to one PRB and the other part of the VRB pair is mapped to another PRB which is at a predefined gap distance (which causes the inter-slot hopping) in order to achieve frequency diversity gain. This mechanism is especially important for small resource block allocations due to inherently less frequency diversity.

The distributed VRBs are mapped to PRBs such that consecutive VRBs are not mapped to sequential PRBs in the frequency-domain. A single VRB pair may be further split in the frequency-domain. Allocation of distributed resources includes a mapping from logically consecutive VRB pairs to physically non-consecutive PRB pairs to provide frequency diversity among VRB pairs. This is performed by a virtual block-interleaver or permutation scheme which operates on resource block pairs. The allocation of distributed resources further includes splitting of each resource-block pair such that the two resource blocks of the resource-block pair are transmitted with a certain frequency gap which provides additional frequency diversity for a single VRB pair. The latter can be interpreted as the introduction of frequency hopping on a slot basis. We will see later in this section that the associated PDCCH downlink control information (DCI) indicates whether allocated VRBs are of a localized or distributed type in the case of type 2 resource allocations. Thus, it is possible to dynamically switch between distributed and localized allocation, and also mix distributed and localized transmission for different terminals within the same subframe. The exact size of the frequency gap in Figure 9.32 depends on the overall downlink system bandwidth as given in Table 9.6. The gaps have been chosen based on the following criteria: (1) the gap size should be in the order of half of the total bandwidth (in terms of the number of resource blocks) in order to provide sufficient frequency diversity gain; and (2) the gap size should be an integer multiple of P^2, where parameter P is the size of a RBG for resource allocation types 0 and 1. The reason for this constraint is to ensure that distributed transmission as described above and transmissions based on downlink allocation types 0 and 1 can coexist on the same subframe. Due to the constraint that the gap size should be an integer multiple of P^2, the gap size N_{gap} may deviate from precisely half of the system bandwidth when the number of resource blocks is an odd number. In those cases, only a fraction of available resource blocks within the system bandwidth can be used for distributed transmission. An example of distributed resource allocation is shown in Figure 9.32.

The permissible values of gap size N_{gap} are given in Table 9.6. Note that for $6 \leq N_{RB}^{DL} \leq 49$, only one gap value $N_{gap} = N_{gap}^{(1)}$ is defined, whereas for $50 \leq N_{RB}^{DL} \leq 110$, two gap values $N_{gap}^{(1)}$ and $N_{gap}^{(2)}$ are defined where the choice of the gap value is signaled as part of the downlink scheduling assignment. VRBs of

Samsung Ex. 1005, Page 488 of 1147

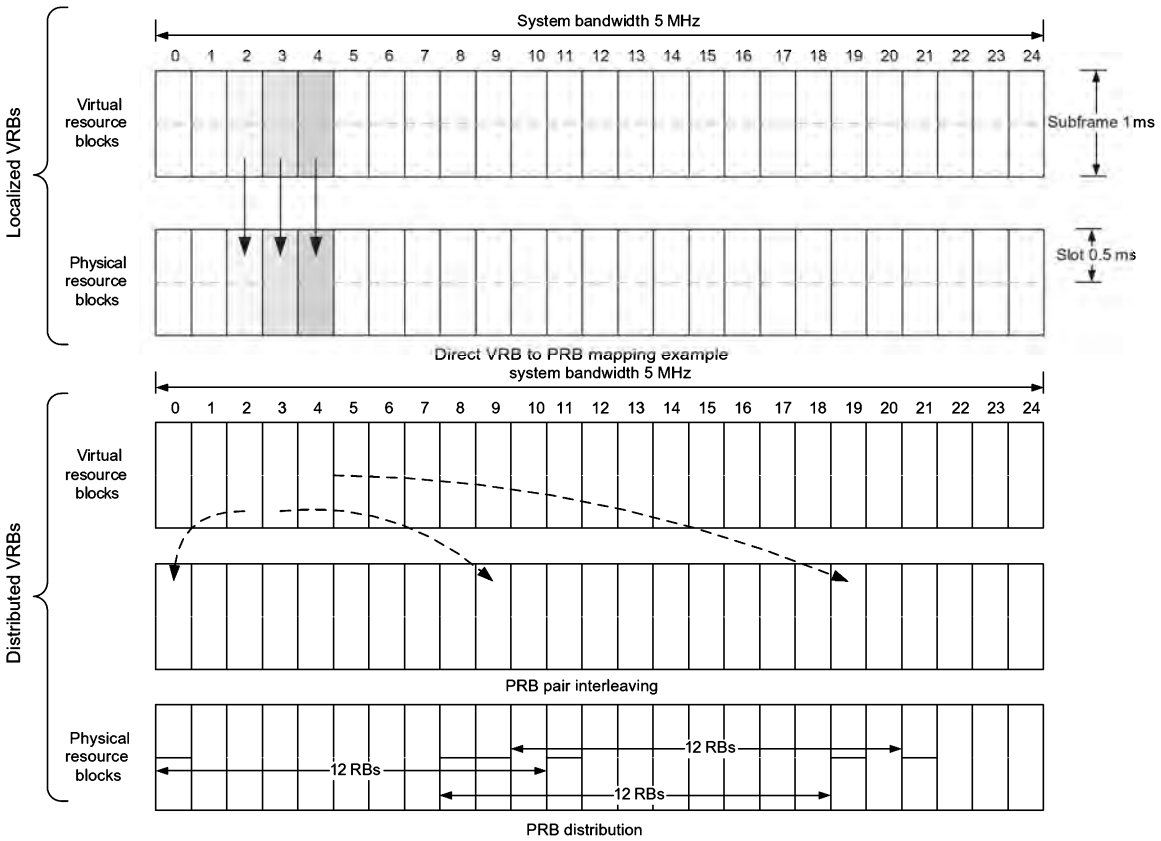


FIGURE 9.32

Illustration of localized and distributed VRB to PRB mapping ($N_{\text{VRB}}^{\text{DL}} = N_{\text{RB}}^{\text{DL}} = 25$) [15].

Table 9.6 Distributed VRB Gap Size as a Function of System Bandwidth [3,5]

Total Number of Resource Blocks N_{RB}^{DL}	Gap Size N_{gap}		Resource Block Group Size P
	$N_{gap}^{(1)}$	$N_{gap}^{(2)}$	
6–10	$\lceil N_{RB}^{DL}/2 \rceil$	–	1
11	4	–	2
12–19	8	–	2
20–26	12	–	2
27–44	18	–	3
45–49	27	–	3
50–63	27	9	3
64–79	32	16	4
80–110	48	16	4

distributed type are numbered from 0 to $N_{VRB}^{DL} - 1$, where $N_{VRB}^{DL} = 2\min(N_{gap}, N_{RB}^{DL} - N_{gap}^{(1)})$ or $N_{VRB}^{DL} = 2N_{gap}^{(2)} \left\lfloor N_{RB}^{DL}/2N_{gap}^{(2)} \right\rfloor$ depending on the value of N_{gap}.

Figure 9.33 illustrates the interleaving table for a 10 MHz channel bandwidth when using $N_{gap}^{(1)}$. The VRB indices are written to the table row-by-row and are sequentially read off the table column-by-column. The nulls in the table are ignored and do not generate outputs. When using $N_{gap}^{(1)}$, the VRBs associated with the first half of the output indices from the interleaving table are mapped to the lower set of PRBs starting from the first PRB. The VRBs corresponding to the second half of the output indices from the interleaving table are then mapped to the upper set of PRBs ending at the last PRB. This mapping is applied to the first slot of the subframe. The mapping is cyclically shifted by the gap value $N_{gap}^{(1)}$ for the second slot of the subframe [13]. As shown in Figure 9.34, when using $N_{gap}^{(2)}$ in a 10 MHz bandwidth, two interleaving tables each with six rows are used to define the mapping from VRBs to PRBs. There are some null entries that are inserted in each table [3].

To better understand the difference between the localized and distributed allocations, consider the example shown in Figure 9.35. The frequency response of a multipath fading channel varies with time and frequency due to path loss, shadowing, and user mobility effects. Let us assume that user A and user B are two users with different channel conditions. The base station receives channel quality reports (typically in the form of CINR or SINR measurements) from the mobile stations in the cell. The base station scheduler may adopt either of the following allocation schemes depending on the channel condition reports and other considerations that will be discussed later in this chapter. One allocation strategy is to allocate user A and user B in the sub-channels where the corresponding SINR is

Samsung Ex. 1005, Page 490 of 1147

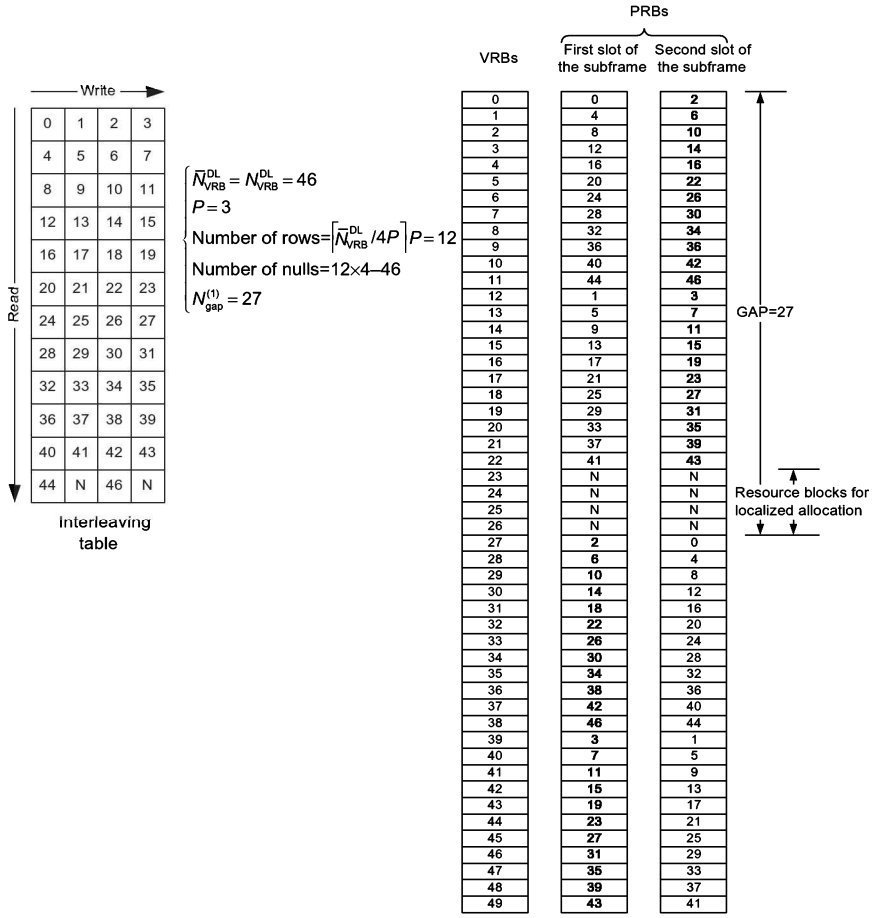


FIGURE 9.33

Illustration of distributed VRB to PRB mapping procedure in 10 MHz using $N_{gap}^{(1)}$ [13].

the best for that user. This is when a frequency-selective scheduling gain can be achieved through allocation of a group of physically adjacent sub-carriers in the sub-channels with relatively strong SINR for that user. Another strategy is to permute the sub-carriers over the entire channel, form a logical group of distributed sub-carriers and assign them to a user. In this case, frequency diversity gain can be achieved through use of distributed resource units for this user.

9.8.2 Resource allocation

One of the main design objectives for signaling of resource allocations, in terms of a set of resource blocks in each subframe, to the active UEs in the cell is to find a

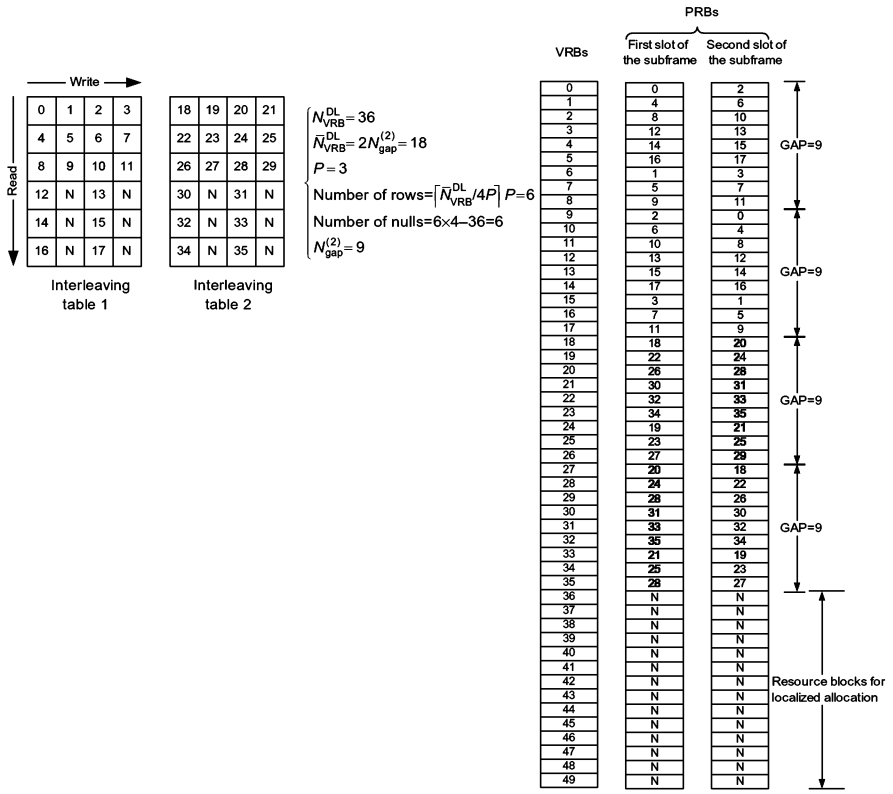


FIGURE 9.34

Illustration of distributed VRB to PRB mapping procedure in 10 MHz using $N_{gap}^{(2)}$ [13].

good trade-off between flexibility and signaling overhead. Indications of localized/distributed resource allocations to different UEs are transmitted via PDCCH. While the exact use of PDCCH depends on the scheduling algorithms implemented in the eNB, it is nevertheless possible to outline some general principles of this operation. The resource allocation field in downlink control information (carried via PDCCH or ePDCCH) is interpreted by the UE depending on the DCI format. A resource allocation field in each DCI includes two parts, a resource allocation header field and information pertaining to the actual resource block assignment. The DCI formats 1, 2, 2A, 2B, 2C and 2D with type 0, and DCI formats 1, 2, 2A, 2B, 2C and 2D with type 1 resource allocation (carried via PDCCH) have the same format and are distinguished from each other via the single bit resource allocation header field which exists depending on the downlink system bandwidth. The PDCCH with DCI format 1A, 1B, 1C, and 1D support type 2 resource allocation, while PDCCH with DCI format 1, 2, 2A, 2B, 2C and 2D support type 0 or type 1

Samsung Ex. 1005, Page 492 of 1147

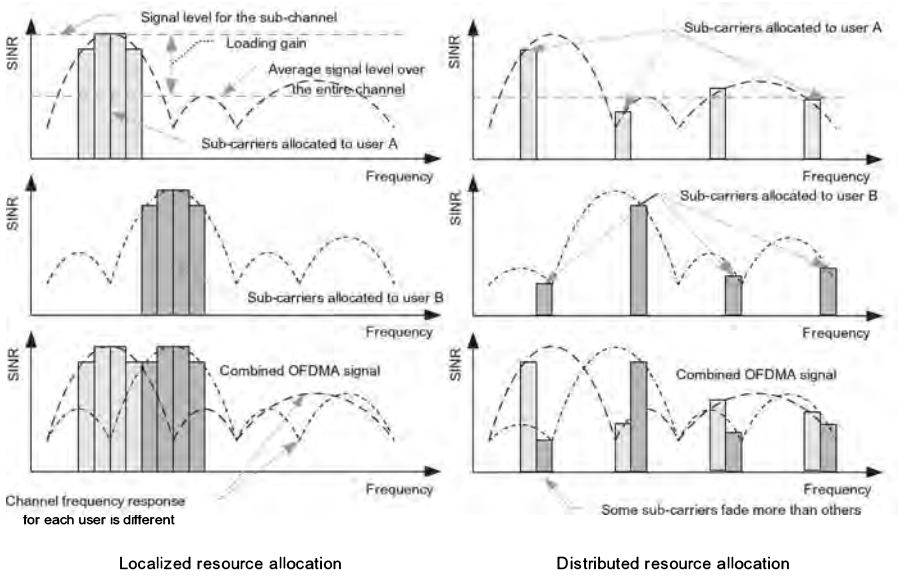


FIGURE 9.35

Comparison of localized and distributed allocation schemes (example) [18].

resource allocation. The DCI formats with a type 2 resource allocation do not have a resource allocation header field [5]. LTE/LTE-Advanced specify three resource allocation schemes in the downlink, where their usage depends on the number resources assigned to the UE, and whether a localized or distributed allocation must be used for the UE considering the channel conditions that the UE is experiencing. In LTE Rel-11, the DCI formats 1, 2, 2A, 2B, 2C and 2D with type 0 and DCI formats 1, 2, 2A, 2B, 2C and 2D with type 1 resource allocation (carried via ePDCCH) have the same format and are distinguished from each other via the single-bit resource allocation header field which exists depending on the downlink system bandwidth. The ePDCCH with DCI format 1A, 1B, and 1D have a type 2 resource allocation.

9.8.2.1 Resource allocation type 0

In resource allocation type 0, the resource block assignment information includes a bitmap indicating the RBGs that are allocated to the scheduled UE where an RBG is a set of consecutive VRBs of localized type. Note that in this case, the granularity of resource assignment is an RBG. The RBG size P is a function of the system bandwidth as given in Table 9.6. The total number of RBGs is given by $N_{\text{RBG}} = \lceil N_{\text{RB}}^{\text{DL}}/P \rceil$. The bitmap is of size N_{RBG} bits and each bit corresponds to a unique RBG in the set. The RBGs are indexed in the order of increasing

frequency and non-increasing RBG sizes starting at the lowest frequency (i.e., RBG_0 to $RBG_{N_{RBG}-1}$ are mapped to the most-significant-bit and least-significant-bit of the bitmap, respectively). The RBG is allocated to a UE, if the corresponding bit value in the bitmap is set to "1." Therefore, each bit in the bitmap will select a small contiguous group of resource blocks whose size depends on the bandwidth. The maximum number of resource blocks assigned using type 0 allocation is the entire bandwidth. As an example, consider a total of 50 resource blocks (i.e., 10 MHz system bandwidth), the number of bits in the bitmap is 17. Each bit in the 17-bit bitmap selects a group of three consecutive resource blocks, with the exception of the last group which will only contain two resource blocks, as shown in Figure 9.36.

9.8.2.2 Resource allocation type 1

In resource allocation type 1, a bitmap indicates PRBs inside a selected RBG subset $0 \leq p < P$. The maximum number of resource blocks assignable through type 1 resource allocation is a subset of the entire bandwidth, i.e., a type 1 allocation, even with all bits in the bitmap set to "1" does not span the entire bandwidth. Each bit in the bitmap will select a single resource block from a set of contiguous group of resources whose size and separation depend on the total bandwidth. In resource allocation type 1, the resource block assignment signaling consists of three parts: (1) select a RBG subset; (2) whether to apply an offset when interpreting the bitmap using a shift flag; and (3) the bitmap that indicates to the UE the specific PRBs inside the selected RBG subset.

The size of the bitmap used to address VRBs in a selected RBG subset is given by $N_{RB}^{(1)} = \lceil N_{RB}^{DL}/P \rceil - \lceil \log_2(P) \rceil - 1$. The addressable VRBs within a selected RBG subset start from an offset $n_{shift}(p)$ to the smallest VRB number within the selected RBG subset. The offset is interpreted in terms of the number of VRBs, and is given by $n_{shift}(p) = N_{RB}^{RBG_{subset}}(p) - N_{RB}^{(1)}$ (if the value of the shift bit is non-zero) where the parameter $N_{RB}^{RBG_{subset}}(p)$ denotes the number of VRBs in RBG subset p and is given as follows [5]:

$$N_{RB}^{RBG_{subset}}(p) = \begin{cases} P\left(\left\lfloor \frac{N_{RB}^{DL}-1}{P^2} \right\rfloor + 1\right) & p < \left\lfloor \frac{N_{RB}^{DL}-1}{P} \right\rfloor \bmod P \\ P\left\lfloor \frac{N_{RB}^{DL}-1}{P^2} \right\rfloor + (N_{RB}^{DL}-1) \bmod P + 1 & p = \left\lfloor \frac{N_{RB}^{DL}-1}{P} \right\rfloor \bmod P \\ P\left\lfloor \frac{N_{RB}^{DL}-1}{P^2} \right\rfloor & p > \left\lfloor \frac{N_{RB}^{DL}-1}{P} \right\rfloor \bmod P \end{cases} \tag{9.23}$$

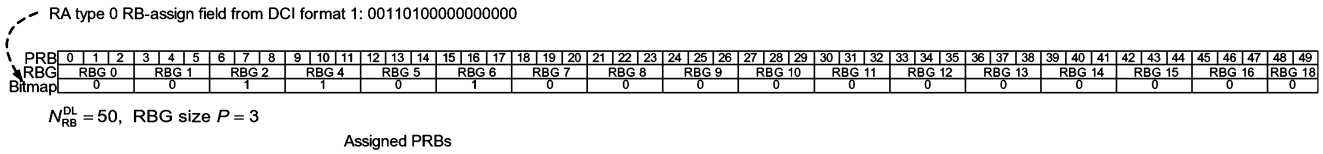


FIGURE 9.36

Example resource allocation type 0 [3,5].

When RBG subset p is addressed, the ith bit ($i = 0, 1, \ldots, N_{RB}^{(1)} - 1$) in the bitmap field corresponds to VRB number $n_{VRB}^{RBG_{subset}}(p) = \lfloor [i + n_{shift}(p)]/P \rfloor P^2 + pP + (i + n_{shift}(p)) \bmod P$. In comparison to resource allocation type 0, the bitmap size for type 1 is always shorter by $\lceil \log_2 P \rceil + 1$ number of bits. As an example, consider a total of 50 resource blocks, the number of bits in the bitmap is 14 (i.e., three bits shorter relative to type 0). Each bit in the 14-bit bitmap will select an individual resource block inside the selected subset, as demonstrated in Figure 9.37.

9.8.2.3 Resource allocation type 2

In resource allocation type 2, the resource block assignment information comprises a set of contiguously allocated localized VRBs or distributed VRBs. When the resource allocation is signaled in PDCCH/ePDCCH with DCI format 1A, 1B, or 1D, a 1-bit flag indicates whether localized VRBs or distributed VRBs are assigned (value "0" indicates localized and value "1" indicates distributed VRB assignment), while distributed VRBs are always assigned in the case of resource allocation signaled with DCI format 1C. A localized VRB assigned to a UE can vary from a single VRB up to a maximum number of VRBs over the system bandwidth. For DCI format 1A, the distributed VRBs allocated to a UE may vary from a single VRB up to N_{VRB}^{DL}, if the DCI CRC bits are scrambled with P-RNTI, RA-RNTI, or SI-RNTI. With DCI format 1B, 1D, and the CRC scrambled with C-RNTI, or with DCI format 1A and the CRC scrambled with C-RNTI, SPS C-RNTI or temporary C-RNTI, distributed VRB allocations allocated to a UE vary from a single VRB up to N_{VRB}^{DL} VRBs, if $6 \le N_{RB}^{DL} < 50$, and vary from a single VRB up to 16 if $50 \le N_{RB}^{DL} < 110$. Using DCI format 1C, the number of distributed VRB allocations for a UE may range from N_{RB}^{step} up to $\lfloor N_{VRB}^{DL}/N_{RB}^{step} \rfloor N_{RB}^{step}$ with an increment of N_{RB}^{step}, where $N_{RB}^{step} = 2$ if $6 \le N_{RB}^{DL} < 50$, and $N_{RB}^{step} = 4$ if $50 \le N_{RB}^{DL} < 110$ [5].

For PDCCH/ePDCCH with DCI format 1A, 1B, or 1D, the type 2 resource allocation field consists of a parameter referred to as resource indication value (RIV) which corresponds to the starting resource block index RB_{start} and the length of the allocation in terms of the number of virtually contiguous resource blocks L_{RB}. The RIV is defined by $RIV = N_{RB}^{DL}(L_{RB} - 1) + RB_{start}$ if $(L_{RB} - 1) \le \lfloor N_{RB}^{DL}/2 \rfloor$; otherwise $RIV = N_{RB}^{DL}(N_{RB}^{DL} - L_{RB} + 1) + (N_{RB}^{DL} - 1 - RB_{start})$, where $1 \le L_{RB} < N_{VRB}^{DL} - RB_{start}$. For DCI format 1C; however, the RIV in type 2 resource allocation field corresponds to starting resource block index $RB_{start} = \{0, N_{RB}^{step}, 2N_{RB}^{step}, \ldots, (\lfloor N_{VRB}^{DL}/N_{RB}^{step} \rfloor - 1)N_{RB}^{step}\}$ and the length in terms of virtually contiguous resource blocks $L_{RB} = \{N_{RB}^{step}, 2N_{RB}^{step}, \ldots, \lfloor N_{VRB}^{DL}/N_{RB}^{step} \rfloor N_{RB}^{step}\}$. In other words, in type 2 resource allocation, PRBs are not directly allocated. The type 2 allocation supports both localized and distributed VRB allocation which are distinguished by a flag. The information regarding the starting point of VRB and the length in terms of contiguously allocated VRB can be derived from RIV signaled within the DCI.

Samsung Ex. 1005, Page 496 of 1147

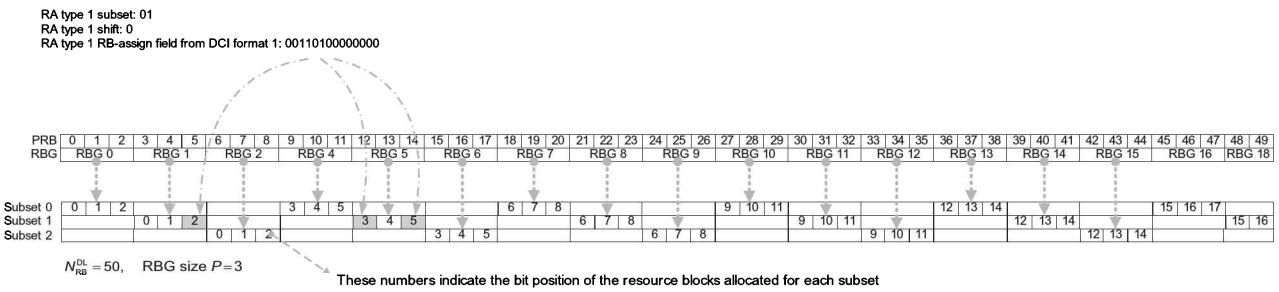


FIGURE 9.37

Example resource allocation type 1 [158].

As an example, consider a total of 50 resource blocks. Let us assume that the eNB is going to assign six localized VRBs starting from resource block 9 in the frequency-domain. Using the above equations for DCI format 1A, the RIV value is calculated as RIV = 259. This RIV is signaled in DCI format 1A, and the UE can unambiguously derive the starting resource block and the number of allocated resource blocks from RIV, as shown in Figure 9.38.

9.8.2.4 PRB bundling

When a UE is configured for operation in transmission mode 9, it may assume that the precoding granularity is multiple resource blocks in the frequency-domain when configured for PMI/RI reporting. In a given serving cell, if a UE is configured in transmission mode 10 and if PMI/RI is reported for all configured CSI processes, then the UE may assume that precoding granularity is multiple resource blocks; otherwise, the precoding granularity is going to be one resource block in the frequency-domain. Fixed bandwidth-dependent precoding resource block groups (PRGs) P' partition the available resource blocks and each PRG consists of a number of consecutive PRBs. If $N_{RB}^{DL} \bmod P' > 0$, then the size of one of the PRGs is adjusted to $N_{RB}^{DL} - P'\lfloor N_{RB}^{DL}/P' \rfloor$. The PRG size is in descending order starting at the lowest frequency index. The UE may assume that the same precoder is applied to all scheduled PRGs within the bundle. The size of the PRG depends on the system bandwidth, and is given in Table 9.7.

9.8.2.5 Resource element group

Resource element groups (REGs) are the basic unit of resource allocation for PCFICH, PHICH, and PDCCH that are represented by an index pair (k', l'), where k' is the sub-carrier index of the resource element within the REG with the lowest sub-carrier index, and l' is the OFDM symbol index of the REG. In other words, resource-element groups are used for defining the mapping of control channels to resource elements. The mapping of a symbol-quadruplet $\{a(i), a(i+1), a(i+2), a(i+3)\}$ to a resource-element group is defined such that elements $a(i)$ are mapped to resource elements (k, l) of the resource-element group not used for cell-specific reference signals (CRSs) in increasing order of indices k and l, as shown in Figure 9.39.

A REG always contains four consecutive resource elements which are not occupied by a CRS on any antenna port in use. All resource elements within a resource block in the first three OFDM symbols are allocated to REGs. Therefore, the number of resource elements within each REG and the number of REGs within an OFDM symbol are affected by the number of CRSs present on all antenna ports. The number and location of cell-specific reference signals are dependent on the number of antenna ports and the type of cyclic prefix used. Each antenna port has a unique CRS associated with it. Considering the dependency of REG arrangement on CRS configuration, the REG structure for one, two, or four antenna port

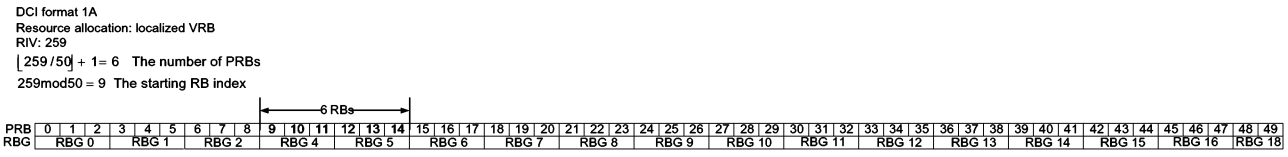


FIGURE 9.38

Example resource allocation type 2 [158].

Table 9.7 Precoding Resource Block Group Size [5]

Number of Resource Blocks N_{RB}^{DL}	Precoding Resource Block Group Size P'
≤10	1
11–26	2
27–63	3
64–110	2

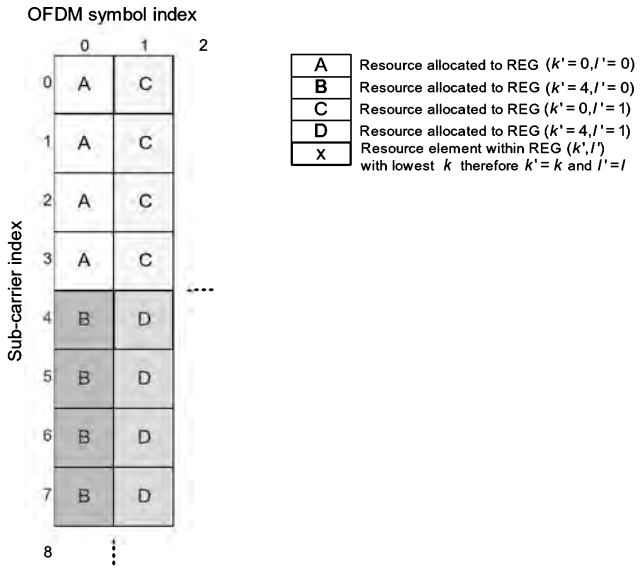


FIGURE 9.39

Resource element group structure [3,160].

configuration is different. However, the REG arrangement for each resource block within a subframe and for each antenna port is identical.

When antenna port 0 or antenna ports 0 and 1 are used, it is assumed that the CRSs are present on both antenna ports 0 and 1. This leads to an REG arrangement for each resource block on the first slot of each subframe, as shown in Figure 9.40. The CRSs are present within the first OFDM symbol. Since four resource elements not including CRSs are required to form a REG, the 12 resource elements over the first symbol in a resource block are divided into two REGs, each containing six resource elements (including two CRSs). On the

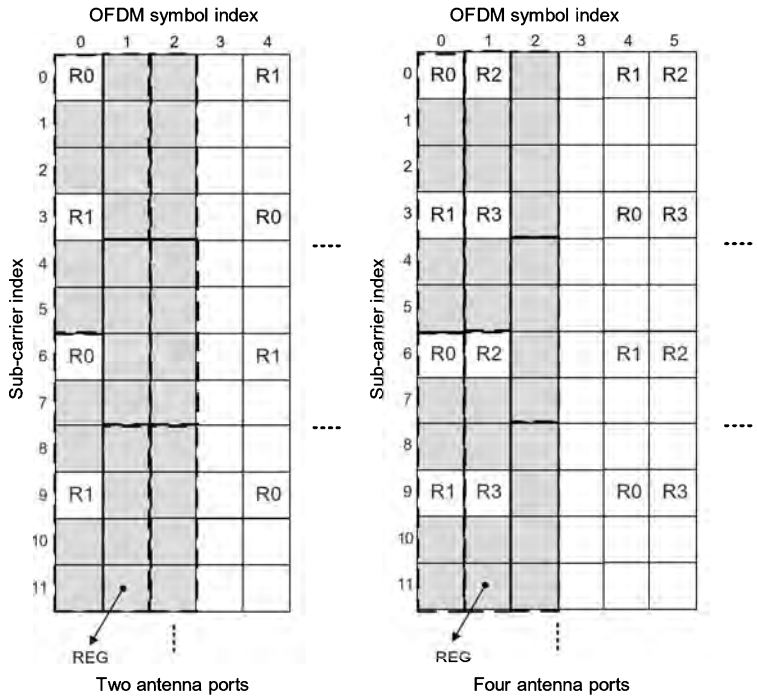


FIGURE 9.40

Resource element group structure with two and four antenna ports [3,160].

second and third OFDM symbols, no CRS is present; therefore; the 12 resource elements over each OFDM symbol are divided into three REGs, each containing four resource elements.

The REG organization in each resource block for four antenna port configuration is illustrated in Figure 9.40. The REG allocation within the first and second OFDM symbols is the same as one or two antenna port configuration. Four CRSs are present on the second OFDM symbol; therefore, eight resource elements are available for the mapping of control information. The 12 reference elements are divided into two REGs, each containing six resource elements. The third OFDM symbol contains no reference signals, so three REGs are available. An extended cyclic prefix subframe contains 12 OFDM symbols as opposed to 14 for the normal cyclic prefix. Since the number of CRSs in a normal or extended cyclic prefix subframe is the same, the limited number of OFDM symbols in an extended cyclic prefix subframe requires the OFDM symbol spacing of the CRSs to be reduced compared to a normal cyclic prefix. This reduction in spacing causes CRSs to be present within the fourth OFDM symbol of an extended cyclic prefix subframe, whereas in a normal cyclic prefix subframe no

CRSs are present on the fourth OFDM symbol of the subframe. The number of cell-specific reference symbols within the first three OFDM symbols is identical for normal or extended cyclic prefix, thus the REG configurations are identical.

9.9 Time and frequency synchronization

Synchronization signals are transmitted in every radio frame. The synchronization signals are used for time and frequency synchronization during initialization. As part of the system acquisition process, the UE synchronizes to the OFDM symbol, slot, subframe, half-frame, and radio frame (in that order) using a synchronization signal. Two synchronization signals are defined for system acquisition in E-UTRA: (1) a primary synchronization signal is used to obtain slot, subframe, and half-frame boundary. It also provides cell identity within the cell identity group. (2) A secondary synchronization signal is used to obtain the radio frame boundary. It also enables the UE to determine the cell identity group, which may range from 0 to 167. The synchronization signals are transmitted in the 0th and 5th subframes of every radio frame. In the frequency-domain, the synchronization signals occupy the center 1.08 MHz (equivalent to six resource blocks) of the radio channel (Figure 9.41).

Figure 9.41 shows the location of the primary and secondary synchronization signals in a radio frame. In the time-domain, primary and secondary synchronization signals are transmitted in the 0th and 10th slot (1st slot within 0th and 5th subframe) of every radio frame. Within these slots, the primary and secondary synchronization signals are transmitted in the last two symbols. The primary synchronization signal is transmitted in the last symbol of the 0th and 10th slot. The secondary synchronization signal is transmitted in one before the last symbol of the 0th and 10th slot. The PBCH is mapped to the first four symbols in slot 1 of the frame (just after the primary and secondary synchronization signals). In the frequency-domain, the center 1.08 MHz of the radio channel is reserved for primary and secondary synchronization signals. This is equivalent to six resource blocks or 72 sub-carriers. However, the primary and secondary synchronization signals only require 62 sub-carriers around the DC sub-carrier. The remaining sub-carriers (10 sub-carriers) in these resource blocks are not used for transmission. They provide a guard band between primary and secondary synchronization signal transmissions and other data transmissions.

There are 504 unique physical layer cell identities. The physical layer cell identities are grouped into 168 unique physical layer cell-identity groups, each group contains three unique identities. The grouping is done such that each physical layer cell identity is part of one and only one physical layer cell-identity group. A physical layer cell identity $N_{\mathrm{ID}}^{\mathrm{cell}} = 3N_{\mathrm{ID}}^{(1)} + N_{\mathrm{ID}}^{(2)}$ is thus uniquely defined by a number $N_{\mathrm{ID}}^{(1)}$ in the range of 0−167, representing the physical layer cell-identity group, and a number $N_{\mathrm{ID}}^{(2)}$ in the range of 0−2, denoting the physical layer identity within the physical layer cell-identity group.

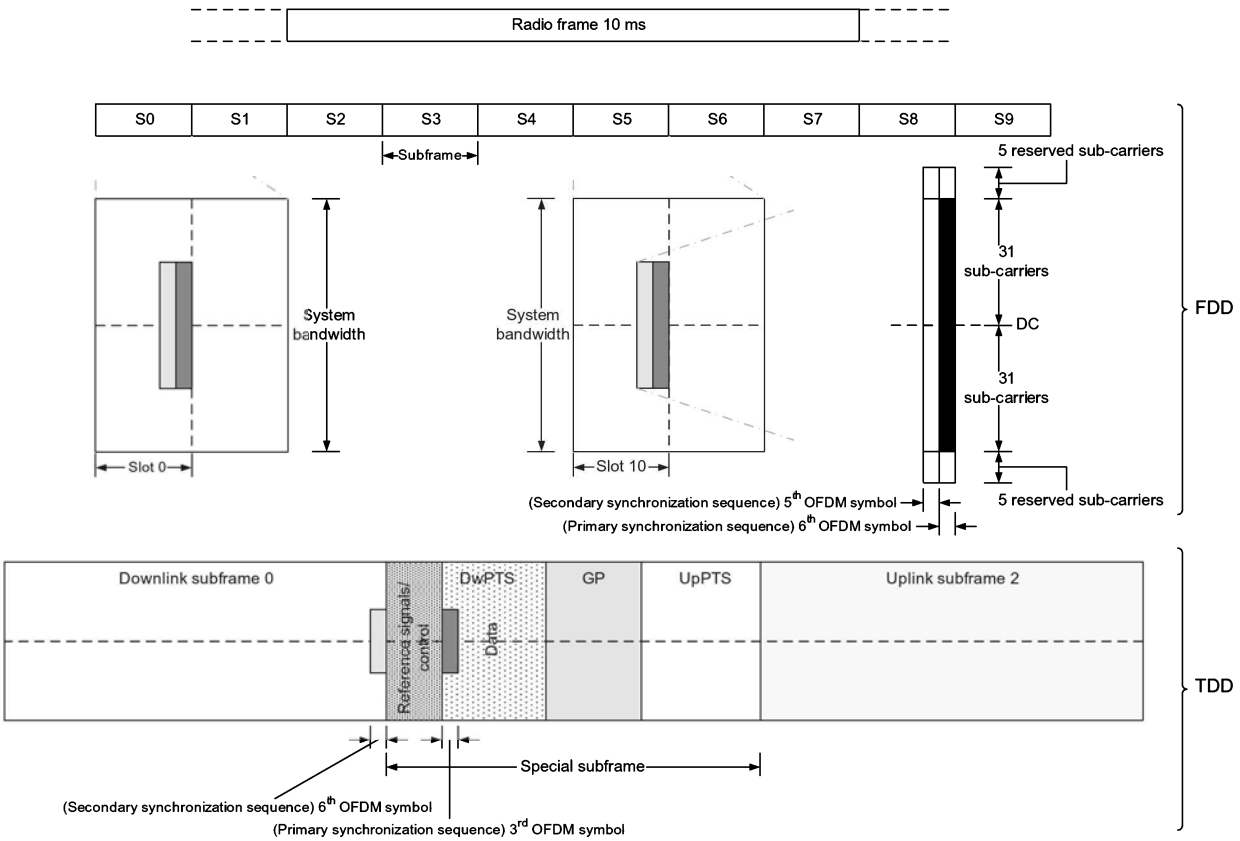


FIGURE 9.41

Location and structure of primary and secondary synchronization signals [3,174].

9.9.1 Primary synchronization signals

The primary synchronization signal is generated from a frequency-domain Zadoff–Chu sequence[9] according to the following equation [3]:

$$d_{\mathrm{u}}(n) = \begin{cases} \mathrm{e}^{-j\frac{\pi un(n+1)}{63}} & n = 0, 1, \ldots, 30 \\ \mathrm{e}^{-j\frac{\pi u(n+1)(n+2)}{63}} & n = 31, 32, \ldots, 61 \end{cases} \tag{9.24}$$

where the Zadoff–Chu root sequence index $u = 25, 29, 34$ if $N_{\mathrm{ID}}^{(2)} = 0, 1, 2$, respectively. The mapping to resource elements depends on the frame structure. The sequence $d(n)$ is mapped to the resource elements according to $a_{\mathrm{kl}} = d(n), n = 0, \ldots, 61$ and $k = n - 31 + 6N_{\mathrm{RB}}^{\mathrm{DL}}$. As shown in Figure 9.41, for frame structure type 1, the primary synchronization signal is mapped to the last OFDM symbol in slots 0 and 10. For frame structure type 2, the primary synchronization signal is mapped to the third OFDM symbol in subframes 1 and 6. The

[9] A Zadoff–Chu sequence is a complex-valued sequence with constant amplitude property whose cyclically shifted versions exhibit low cross-correlation. Thus, under certain conditions, the cyclically shifted versions of each sequence remain orthogonal to one another. A Zadoff–Chu sequence that has not been shifted is referred to as a root sequence. The uth root Zadoff–Chu sequence of prime length N is defined as follows [172]:

$$x_u(n) \triangleq \begin{cases} \mathrm{e}^{-\mathrm{j}[\pi un(n+1)]/N} & 0 \le n < N-1 \quad (N \text{ is an odd integer}) \\ \mathrm{e}^{-\mathrm{j}\pi un^2/N} & 0 \le n < N-1 \quad (N \text{ is an even integer}) \end{cases}$$

where N is an integer, which denotes the length of the Zadoff–Chu sequence. It is easily verified that $x_u(n)$ is periodic with period N, i.e., $x_u(n) = x_u(n+N), \forall n$. In other words, the sequence index u is prime relative to N. For a fixed value of u, the Zadoff–Chu sequence has an ideal periodic autocorrelation property (i.e., the periodic autocorrelation is zero for all time shifts other than zero). For different values of index u, the Zadoff–Chu sequences are not orthogonal, but rather exhibit low cross-correlation. If the sequence length N is selected as a prime number, there are different sequences with periodic cross-correlation of $1/\sqrt{N}$ between any two sequences regardless of time shift. The Zadoff–Chu sequences are a subset of constant amplitude zero autocorrelation (CAZAC) sequences. The properties of Zadoff–Chu sequences can be summarized as follows [172]:

- They are periodic with period N, if N is a prime number, i.e., $x_u(n+N) = x_u(n)$.
- Given N is a prime number, the DFT of a Zadoff–Chu sequence is another Zadoff–Chu sequence conjugated and time-scaled multiplied by a constant factor, i.e., $X_u[k] = x_u^*(vk)X_u[0]$ where v is the multiplicative inverse of u modulo N. It can be shown that $x_u^*(vk) = x_v^*(k)\mathrm{e}^{\mathrm{j}\pi(1-v)k/N}$.
- The autocorrelation of a prime-length Zadoff–Chu sequence with a cyclically shifted version of itself also yields zero autocorrelation sequence, i.e., it is non-zero only at one instant which corresponds to the cyclic shift zero.
- The cross-correlation between two prime-length Zadoff–Chu sequences, i.e., different u, is constant and equal to $1/\sqrt{N}$.
- The Zadoff–Chu sequences have low PAPR.

resource elements a_{kl} over OFDM symbols used for transmission of the primary synchronization signal where $k = n - 31 + 6N_{RB}^{DL}, n = \{-5, -4, \ldots, -1, 62, 63, \ldots, 66\}$ are reserved as guard sub-carriers.

9.9.2 Secondary synchronization signals

The secondary synchronization signal is generated from interlacing of two length-31 M-sequences[10] $s_0^{(m_0)}(n)$ and $s_1^{(m_1)}(n)$. There are 168 sequences that are generated from cyclic shifts of sequences $s_0^{(m_0)}(n)$ and $s_1^{(m_1)}(n)$. The cyclic shifts are generated based on cell-identity group number, which ranges from 0 to 167. The two sequences are scrambled, with the scrambling sequences $c_0(n)$ and $c_1(n)$ that are obtained based on the primary synchronization signal identity $N_{ID}^{(2)} \in \{0, 1, 2\}$. The second sequence is further scrambled with additional sequence $z_1^{(m_0)}(n)$. The two sequences are interleaved and mapped to 62 sub-carriers around the DC sub-carrier using BPSK modulation. As shown in Figure 9.42, five sub-carriers on either side are not used, providing a guard band between the primary and secondary synchronization sequences and other data-bearing sub-carriers. The secondary synchronization signals are also used to detect the radio frame boundary. To achieve this, a different variation of the sequence is generated by switching $s_0^{(m_0)}(n)$ and $s_1^{(m_1)}(n)$ in slot 10. The switch produces a different sequence so the UE can distinguish between slot 0 and slot 10.

More specifically, the secondary synchronization signal $d(0), d(1), \ldots, d(61)$ is an interleaved concatenation of two length-31 binary sequences as shown in Figure 9.42. The concatenated sequence is scrambled with a scrambling sequence given by the primary synchronization signal. The combination of two length-31 sequences defining the secondary synchronization signal differs between subframe 0 and subframe 5 according to the following rule:

$$d(2n) = \begin{cases} s_0^{(m_0)}(n)c_0(n) & \text{in subframe 0} \\ s_1^{(m_1)}(n)c_0(n) & \text{in subframe 5} \end{cases} \tag{9.25}$$

[10]Maximum length sequences are pseudo-random binary sequences that are generated using maximal linear feedback shift registers. The M-sequences are periodic and reproduce every binary sequence that can be reproduced by the shift registers (i.e., for length-m registers, they produce a sequence of length $2^m - 1$). An M-sequence is spectrally flat with the exception of a near-zero DC term. Since M-sequences are periodic and shift registers cycle through every possible binary value with the exception of the zero vectors, the registers can be initialized to any state with the exception of the zero vectors. A binary polynomial over $GF(2)$ can be associated with the linear feedback shift register. The degree of polynomial is equal to the length of the shift register and the coefficients that are either 0 or 1 correspond to the taps of the register [173].

Samsung Ex. 1005, Page 505 of 1147

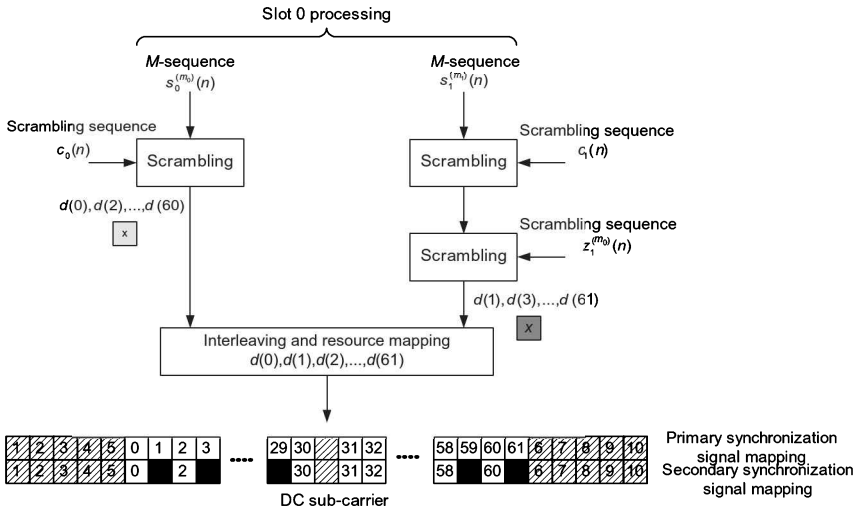


FIGURE 9.42

Secondary synchronization signal generation and mapping.

and:

$$d(2n+1) = \begin{cases} s_1^{(m_1)}(n)c_1(n)z_1^{(m_0)}(n) & \text{in subframe 0} \\ s_0^{(m_0)}(n)c_1(n)z_1^{(m_1)}(n) & \text{in subframe 5} \end{cases} \tag{9.26}$$

where $0 \leq n \leq 30$ and indices m_0 and m_1 are derived from the physical layer cell-identity group $N_{\text{ID}}^{(1)}$. The two sequences $s_0^{(m_0)}(n)$ and $s_1^{(m_1)}(n)$ are derived from two different cyclic shifts of an M-sequence $\tilde{s}(n)$. The two scrambling sequences $c_0(n)$ and $c_1(n)$ depend on the primary synchronization signal and are characterized by two different cyclic shifts of the M-sequence $\tilde{c}(n)$. The scrambling sequences $z_1^{(m_0)}(n)$ and $z_1^{(m_1)}(n)$ are defined by a cyclic shift of the M-sequence $\tilde{z}(n)$. The mapping of the sequence to resource elements depends on the frame structure. In a subframe for frame structure type 1 and in a half-frame for frame structure type 2, the same antenna port used for transmission of the primary synchronization signal is used for transmission of the secondary synchronization signal. The sequence $d(n)$ is mapped to resource elements according to $a_{\text{kl}} = d(n)$, $\forall n = 0, \ldots, 61$ and $k = n - 31 + 6N_{\text{RB}}^{\text{DL}}$, where $l = N_{\text{s}}^{\text{DL}} - 2$ in slots 0 and 10 for frame structure type 1 and $l = N_{\text{s}}^{\text{DL}} - 1$ in slots 1 and 11 for frame structure type 2 [3].

The primary synchronization signals are detected using non-coherent detection, whereas the secondary synchronization signals can be detected using either coherent or non-coherent detection schemes. Figure 9.43 shows an example algorithm for coherent detection of the secondary synchronization signals. The channel compensation block is used for coherent detection. The received secondary synchronization signals in the frequency-domain can be expressed as

Samsung Ex. 1005, Page 506 of 1147

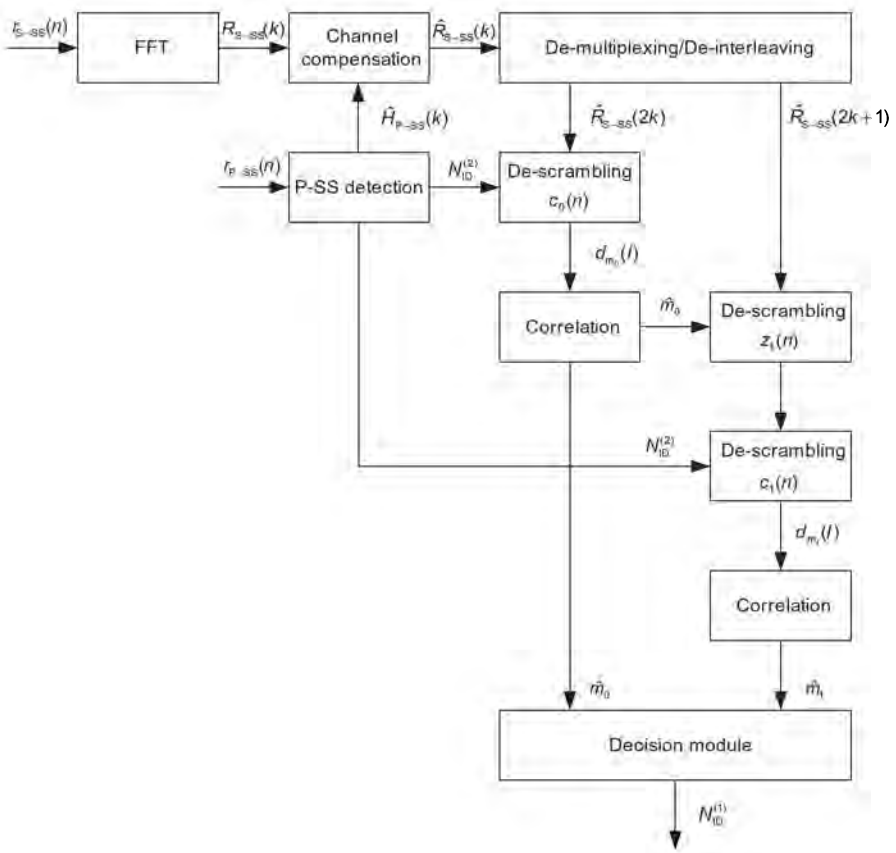


FIGURE 9.43

Example algorithm for coherent detection of secondary synchronization signal [174].

$r_{S-SS}(k) = H(k)d_{S-SS}(k) + w(k)$, where $d_{S-SS}(k)$ is the secondary synchronization signal in the frequency-domain, $H(k)$ is channel frequency response during secondary synchronization signal transmission, and $w(k)$ is the additive white Gaussian noise. For coherent detection, the UE estimates the channel frequency response $\hat{H}(k)$ by using the received primary synchronization sequence. From the UE perspective, the secondary synchronization signal detection is performed after the primary synchronization signal detection; thus the channel can be assumed to be known based on the primary synchronization signal sequence. In the frequency-domain, channel compensated secondary synchronization signal is given by $\hat{R}_{S-SS}(k) = \hat{H}(k)R_{S-SS}(k)$. After channel compensation, the de-interleaved and de-scrambled signal can be expressed as $d_{m_0}(l) = \hat{R}_{S-SS}(2k)c_0(k), d_{m_1}(k) = \hat{R}_{S-SS}(2k+1)c_1(k)z_1^{(\hat{m}_0)}(k)$. The output of coherent detection is the cross-correlation

between the de-scrambled signal and the cyclic shift of a secondary synchronization sequence that can be represented as follows:

$$\hat{m}_0 = \arg \max_i \left| \sum_{l=0}^{30} d_{m_0}(l) s_i(l) \right|^2$$
$$\hat{m}_1 = \arg \max_i \left| \sum_{l=0}^{30} d_{m_1}(l) s_i(l) \right|^2 \quad (9.27)$$

where $\hat{m}_0$ and $\hat{m}_1$ are the ith secondary synchronization signal sequence. Once $\hat{m}_0$ and $\hat{m}_1$ are identified, the frame timing and cell identity group $N_{\mathrm{ID}}^{(1)}$ can be detected (Figure 9.44).

The E-UTRA supports normal and extended cyclic prefix types for various deployment scenarios. The timing of the secondary synchronization signals will change depending on the cyclic prefix type, as shown in Figure 9.44. Prior to successful detection of the secondary synchronization signal, the cyclic prefix type is unknown to the UE and it must be blindly detected by examining the secondary synchronization signals at two possible locations in time. A simple algorithm for determining the cyclic prefix type is introduced in this section as an example. Note that E-UTRA specifications do not specify the receiver architecture and detection methods. This algorithm is based on the principle developed by Van de Beek, where the correlation between the cyclic prefix and its repetition part of the data symbol is calculated for two choices of cyclic prefix type [175]. For each cyclic prefix type, the correlation function is normalized by the associated received secondary synchronization signal power, because the range of values over which the summation is calculated varies with the cyclic prefix type. Let $r(n)$ denote the secondary synchronization signal at the receiver in the time-domain, we define [174]:

$$\rho_i = \sum_{n=0}^{L_i-1} r(n+\tau_i) r^*(n+\tau_i+N)$$
$$P_i = \sum_{n=0}^{L_i-1} |r(n+\tau_i+N)|^2 \quad (9.28)$$

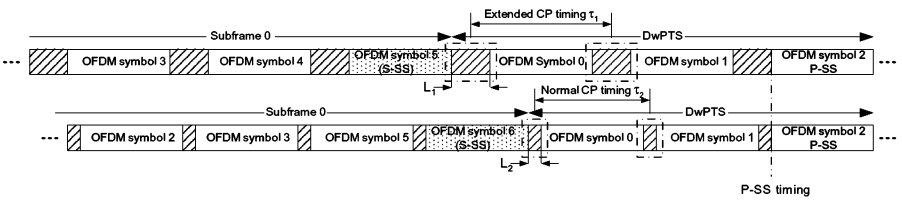


FIGURE 9.44

Illustration of the cyclic prefix detection algorithm [174].

where τ_1 and τ_2 represent the normal and extended cyclic prefix timing, respectively, and L_1 and L_2 denote the corresponding cyclic-prefix lengths. The normalized correlation function $U_i = \mathrm{Re}\{\rho_i\}/P_i$ with the larger value indicates the type of cyclic prefix. Therefore, a normal cyclic prefix is considered, if $U_1 > U_2$ and an extended cyclic prefix is assumed, if $U_2 > U_1$. The main advantage of this algorithm is that a correlation value can be accumulated by using an OFDM symbol within one radio frame in order to improve the detection performance at low SNR. In addition, this algorithm can reduce the computational complexity compared to blind detection using secondary synchronization signal detection at the two possible instances in time.

9.10 Downlink reference signals and channel estimation

To facilitate estimation of the multipath communication channel and reliable, coherent detection of data/control signaling, an OFDM system uses reference signals (or pilot symbols). The pilot symbols provide an estimate of the channel frequency response at the pilot locations over the time–frequency grid. It is possible to estimate the channel at other time–frequency locations using interpolation techniques. In E-UTRA, the CRSs are assigned to specific positions within a subframe, depending on the physical cell identity and the specific antenna port from which they are transmitted. The unique positioning of the pilots ensures that they do not interfere with each other and can be used to provide a reliable estimate of the complex-valued channel gains scaling each resource element within the transmitted grid. Using known pilot symbols to estimate the channel matrix, it is possible to equalize the effects of the channel and to reduce noise and interference effects of the received resource blocks. Downlink spatial multiplexing in LTE Rel-10 was enhanced to support up to eight transmission layers, together with an enhanced reference-signal structure. Relying on cell-specific reference signals for higher order spatial multiplexing is less attractive, as the reference signal overhead is not proportional to the instantaneous transmission rank, but rather to the maximum supported transmission rank. LTE Rel-10, therefore, introduced extensive support of UE-specific reference signals for demodulation of up to eight layers. In addition, feedback of channel-state information (CSI) is based on a separate set of reference signals, known as CSI reference signals (CSI-RSs). The CSI-RSs are relatively sparse in frequency but regularly transmitted from all antennas at the base station, while the UE-specific reference signals are denser in frequency but only transmitted when data is transmitted on the corresponding layer. Separating the reference-signal structures supporting demodulation and channel estimation helps to reduce reference-signal overhead, especially for high degrees of spatial multiplexing, and allows for implementation of various beam-forming schemes. In LTE Rel-8, the reference signal is added to the signal after precoding, one CRS per antenna port. From the received CRS, the UE can estimate the channel matrix and determine how the radio channel affected the signal. Using the estimated channel matrix

together with the knowledge about the code-book used for precoding, the UE can demodulate the received signal and regenerate the original information. In LTE Rel-10, the demodulation reference signals (DM-RSs) are added to the different data streams before precoding. The knowledge about the reference signal will then provide information about the combined effect of radio channel and precoding, thus prior knowledge about the precoder is not required by the receiver, this case is referred to as non-codebook-based precoding.

The LTE/LTE-Advanced systems make use of reference signals for downlink and uplink channel estimation. The downlink reference signals are used to measure CSI and channel estimation for demodulation. The design of downlink reference signals in LTE Rel-8 primarily relied on a set of cell-specific reference signals that could be used for both purposes. The CRS are wideband reference signals transmitted across the entire downlink system bandwidth in every subframe, and are defined for up to four antenna ports. The presence of cell-specific reference signals in every subframe, while necessary for decoding essential physical channels, results in unreasonably high overhead. To support eight-layer spatial multiplexing in LTE-Advanced, new low duty-cycle wideband reference signals, known as CSI-RSs, were designed for CSI measurements. Furthermore, a new set of UE-specific or DM-RSs were designed to enable channel estimation for demodulation. The LTE Rel-10 DM-RS design efficiently supports high-rank transmissions because it is present only in those PRBs in which higher rank is used [121]. The enhanced PDCCH (ePDCCH) DM-RSs with an identical structure to Rel-10 DM-RSs were added in Rel-11 to support various use cases involving the new FDM-based control channel structure.

LTE/LTE-Advanced defines an antenna port as a logical entity which is distinct from a physical antenna and is associated with a specific set of reference signals such that the channel over which a symbol on the antenna port is conveyed can be distinguished from the channel over which another symbol on the same antenna port is conveyed. There is one resource grid per antenna port. The set of downlink antenna ports supported depends on the reference signal configuration in the cell as given in Table 9.8 [3].

The set of downlink antenna ports supported depends on the reference signal configuration in the cell as follows [3]:

- Cell-specific reference signals support one, two, or four antenna port configurations and are transmitted on antenna ports $p = 0$, $p \in \{0, 1\}$, and $p \in \{0, 1, 2, 3\}$, respectively.
- MBSFN reference signals are exclusively transmitted on antenna port $p = 4$.
- UE-specific or DM-RSs are transmitted on antenna port $p = 5, 7,$ or 8, or one or several of $p \in \{7, 8, 9, 10, 11, 12, 13, 14\}$.
- Positioning reference signals are exclusively transmitted on antenna port $p = 6$.
- CSI reference signals support configuration of one, two, four, or eight antenna ports, and are transmitted on antenna ports $p = 15$, $p = \{15, 16\}$, $p = \{15, 16, 17, 18\}$, and $p = \{15, 16, 17, 18, 19, 20, 21, 22\}$, respectively.

Table 9.8 Downlink Antenna Ports and Their Use Cases [3,13]

3GPP Release	Downlink Antenna Port	Associated Reference Signals	Use Cases
8	0–3	Cell-specific reference signals	Single-layer transmission, transmit diversity, SU-MIMO/MU-MIMO
8	4	MBSFN reference signal	Multimedia broadcast and multicast services
8	5	UE-specific reference signal/ demodulation reference signal	Single-layer beamforming
9	6	Positioning reference signal	Location-based services
9	7,8	UE-specific reference signals/demodulation reference signals	Dual-layer beamforming, MU-MIMO
10	9–14	UE-specific reference signals/demodulation reference signals	Multi-layer beamforming, MU-MIMO
10	15–22	CSI reference signals	CSI measurement and reporting
11	107–110	ePDCCH demodulation reference signals	ePDCCH-based transmission and demodulation

- ePDCCH DM-RSs are transmitted on antenna ports $p \in \{107, 108, 109, 110\}$, and are used to support ePDCCH transmission and demodulation.

Based on the above discussion, six types of downlink reference signals are defined as follows, which will be discussed in detail in the next sections:

- Cell-specific reference signals (CRS)
- MBSFN reference signals
- UE-specific or DM-RSs associated with PDSCH
- Positioning reference signals
- CSI-RSs
- DM-RSs associated with ePDCCH

Note that there is one reference signal transmitted per downlink antenna port.

9.10.1 Reference signal design criteria

In this section, we discuss the theoretical criteria for efficient pilot pattern design for OFDM systems. Pilot-based channel estimation is defined as the use of the channel samples estimated at the pilot tones to reconstruct channel samples at the

remaining data tones. As a result, the pilot pattern design is essentially a conventional sampling rate selection problem in the two-dimensional signal processing space. In order to avoid aliasing during reconstruction of the channel time–frequency function, the pilot tone selection should follow the two-dimensional sampling theorem. When multiple antennas are used, the receiver must estimate the channel impulse response (or the transfer function) from each of the transmit antennas in order to correctly detect the signal. This is achieved through distributing reference signals (or pilot tones) among the transmit antennas.

Let $\Delta f = 1/T_{\mathrm{u}}$ and $T_{\mathrm{symbol}} = T_{\mathrm{g}} + T_{\mathrm{u}}$ denote the sub-carrier spacing and the OFDM symbol duration (inclusive of the guard interval), respectively. Let us further assume that the pilot sub-carriers are transmitted at integer multiples of sub-carrier spacing and OFDM symbol duration in frequency and time directions, respectively (i.e., $f_{\mathrm{p}} = m/T_{\mathrm{u}}$ and $T_{\mathrm{p}} = nT_{\mathrm{symbol}}$, where m and n are integers). The (m, n) pair represents the pilots' separation in terms of sub-carrier spacing and OFDM symbol duration. From the sampling theorem point of view, the channel's two-dimensional delay-Doppler response $h(\tau, \nu)$ can be fully reconstructed, if the two-dimensional transform function $H(t,f)$ is sampled greater than or equal to the Nyquist rate across time and frequency dimensions. Hence, the time-domain sampling rate must be greater than or equal to the channel's maximum Doppler spread, i.e., $T_{\mathrm{p}} \leq 1/\nu_{\max}$ (the sampling rate in time must be less than the coherence time) and the frequency-domain sampling rate must be greater than or equal to the channel's maximum delay spread, i.e., $f_{\mathrm{p}} \leq 1/\tau_{\max}$ (the sampling rate in frequency must be less than coherence bandwidth). Assuming a wide-sense stationary uncorrelated scattering channel model, and assuming the channel to be constant over one OFDM symbol, the frequency response $H(t,f)$ of the L-path channel is given as [23,140]:

$$H(t,f) = \frac{1}{\sqrt{L}} \sum_{l=1}^{L} \mathrm{e}^{\mathrm{j}(\psi_l + 2\pi\nu_l t - 2\pi f \tau_l)} \tag{9.29}$$

where ψ_l, ν_l, and τ_l denote the phase, Doppler frequency, and delay of the lth path, respectively. All these parameters are independent random variables. In general, the pilot signals are over-sampled in order to ensure a good trade-off between performance and overhead. Therefore, the choice of (m, n) depends on the channel's maximum delay spread and maximum Doppler spread, and must satisfy the following equation according to the two-dimensional sampling theorem [23]:

$$\begin{aligned} n &\leq \frac{1}{2T_{\mathrm{symbol}}\nu_{\max}} \\ m &\leq \frac{T_u}{2\tau_{\max}} \end{aligned} \tag{9.30}$$

It should be noted that the pilot density for a regular pattern can be calculated using Eq. (9.30). From Eq. (9.30), it can be seen that for large values of $\nu_{\max}$

(i.e., large Doppler spread, or alternatively small channel coherence time, means that the channel is time-varying), n should be small in order to appropriately track channel time variations. On the other hand, for large values of τ_{max} (i.e., large delay spread, or alternatively small coherence bandwidth, means that the channel is frequency selective), m should be small in order to closely follow channel frequency variation. In a regularly spaced pilot pattern, the pilot symbols are evenly spaced in frequency and in time.

Using the default OFDM symbol duration and the sub-carrier spacing given in Table 9.2 (OFDM parameters $T_{symbol} = 71.36\ \mu s$ and $\Delta f = 1/T_u = 15$ kHz), and assuming that the channel maximum delay spread $\tau_{max} = 5\ \mu s$ for the radio channels under consideration in LTE/LTE-Advanced evaluation methodology, the most appropriate range for the values of (m, n) can be determined. Since an E-UTRA system is required to support vehicle speeds up to 350 km/h, the maximum Doppler spread is estimated as $\nu_{max} \approx 810$ Hz at 2.5 GHz carrier frequency. From Eq. (9.30) and the above design parameters, the following pilot spacing constraints are obtained $n \leq 8, m \leq 6$. Figure 9.45 illustrates a regularly spaced

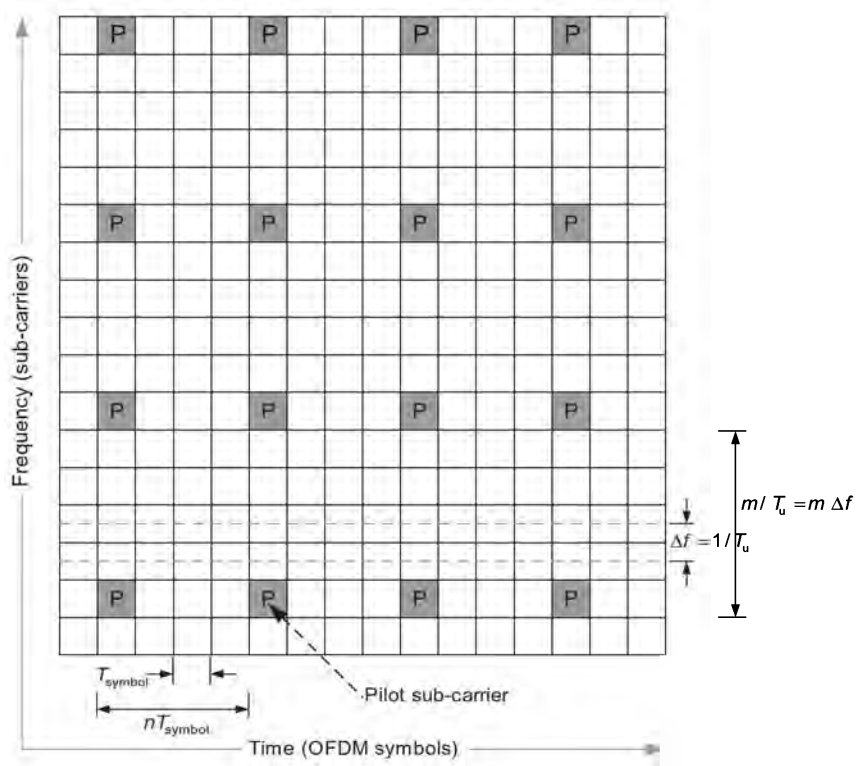


FIGURE 9.45

An example of regular pilot pattern for one spatial layer with $n = 4$ and $m = 5$.

two-dimensional pilot pattern for one spatial stream in the time–frequency domain where $n = 4$ and $m = 5$. The above criteria can be generalized to the case of multi-antenna transmission where each transmit antenna is associated with a group of pilot sub-carriers.

In the case of irregular pilot patterns, the pilot symbols can be irregularly placed in time, in frequency, or in both. The irregular patterns can be designed according to a certain criteria, yielding a specific pattern. Other important considerations in the design of pilot structures for support of multi-antenna OFDM systems include the following:

- Identical pilot pattern for each PRB: Pilot spacing in time and frequency must be in proportion with the resource block size. If the location of the pilot tones within a resource block is not maintained the same across all resource blocks within the system bandwidth, or alternatively the pilot sub-carriers have different positions within each resource block, the filtering and interpolation operations during channel estimation become excessively complex.
- Pilot density: Proper pilot overhead must be considered as a trade-off between accurate channel estimation under various mobility conditions and higher throughput.
- Types of pilots: Pilots are typically classified as common (cell-specific) and dedicated (UE-specific). Each UE in the cell can estimate the channel over the entire bandwidth using the common pilots, while dedicated pilots in the UE assigned resource block can be used for channel estimation. In conventional MIMO schemes, the dedicated pilots are typically adaptively precoded. The dedicated pilots can be used in conjunction with common pilots, or in place of common pilots.
- Pilot power boosting: In general, the power of pilot sub-carriers is boosted relative to data sub-carriers in order to enhance the channel estimation with pilot hopping or shifting. To avoid symbol to symbol power fluctuation when using pilot boosting, it is desirable to place the pilot sub-carriers regularly on every symbol, or to employ power adjustment for data sub-carriers over a symbol.
- Per antenna power balance: When using multiple transmit antennas, it is important to maintain balanced power distribution across antennas. For this purpose, if there are two transmit antennas, each OFDM symbol should contain the same number of pilot tones associated with different antenna ports.

As an example and by taking the above additional criteria into consideration, the structure and link-level performance (i.e., bit error rate versus SNR curves) of several hypothetical common pilot patterns for two transmit antennas and two different multi-antenna transmission schemes are shown in Figure 9.46. In this example, the pilot overhead of type A and type B patterns is 11.11% (i.e., 5.5% per antenna port), and the transmit power per antenna is balanced. The type C

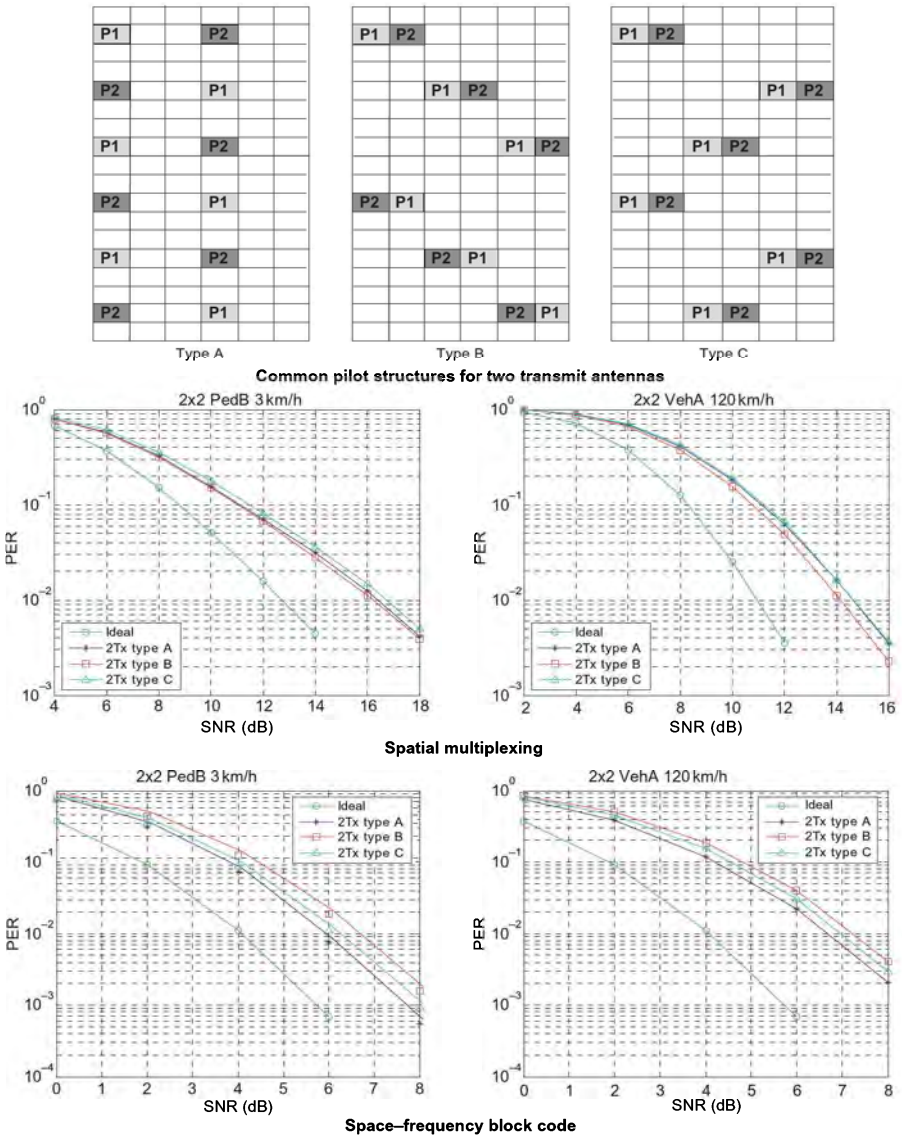


FIGURE 9.46

Channel estimation accuracy of various pilot patterns for two transmit antennas and QPSK-1/2 in different mobility and channel conditions [140].

pilot pattern has unbalanced power across antennas and symbols. The pilot tone powers have been boosted by 3 dB relative to data sub-carriers [140].

The accuracy of channel estimation using common pilots affects the performance of downlink/uplink data and control channel decoding. Figure 9.46 shows the link-level performance of spatial multiplexing with the example pilot patterns, where type B provides the best performance under low and high mobility conditions. The ideal channel estimation is used as a reference. Also shown in Figure 9.46 is the link-level performance of space-frequency block codes (SFBCs) with the example pilot patterns. In this case, type A outperforms the other structures. It can be concluded that type A is the best among three example patterns. Note that the example pilot patterns satisfy the sampling theorem constraint of Eq. (9.30). This example underlines the criteria for the design of robust and high performance pilot structures for OFDM-based cellular system operation in a wide range of mobility and channel conditions. The time/frequency characteristics of a number of typical radio channel models, mobility classes, and the associated time–frequency sampling rate requirements are summarized in Table 9.9.

Table 9.9 Characteristics of Typical Radio Channel Models ($T_u = 66.67$ μs and $T_{symbol} = 71.36$ μs)

Channel Model	RMS Delay Spread (ns) ν_{RMS}	Maximum Delay Spread (μs) τ_{max}	Maximum Sampling Rate in Frequency (Number of Sub-carriers) $\frac{T_u}{2\tau_{max}}$
ITU-T Pedestrian A (PedA)	45	0.41	81
ITU-T Pedestrian B (PedB)	750	3.7	9
ITU-T Vehicular A (VehA)	370.4	2.51	13
ITU-T Vehicular B (VehB)	4000	20	1
Typical Urban 6-Ray Model (TU)	1000	5	6

Vehicular Speed Range (km/h)	Maximum Doppler Spread (Hz) ν_{max}	Coherence Time (ms) T_c	Maximum Sampling Rate in Time (Number of OFDM Symbols) $\frac{1}{2T_{symbol}\nu_{max}}$
0–15	~27.78	15.23	252
15–120	~222.22	1.90	31
120–350	~648.15	0.65	10

9.10.2 Pilot-based channel estimation

While perfect knowledge of the radio channel can be used to find an upper bound for system performance, such knowledge is not available in practice and the channel needs to be frequently estimated. Channel estimation can be performed in various ways, including the use of frequency and/or time correlation properties of the wireless channel, blind or pilot (reference signal) based channel estimation, and adaptive or non-adaptive channel estimation. Non-parametric methods attempt to estimate the frequency response without relying on a specific channel model. In contrast, the parametric estimation methods assume a certain channel model and determine the parameters of this model. Spaced-time and spaced-frequency correlation functions, discussed earlier, are specific properties of a channel that can be incorporated in the estimation method, improving the quality of estimations. Pilot-based estimation methods are the most commonly used methods in OFDM systems which are applicable in systems where the sender transmits some known signals to the receiver. Blind estimation, on the other hand, relies on some properties of the signal and is rarely used in practical OFDM systems. Adaptive channel estimation methods are typically used for a rapid time-varying channel [69–71].

In this section, we describe a generic pilot-based non-adaptive channel estimation method. In a linear time-invariant AWGN channel, the relationship between the transmitted signal $X_k|_{k=0,1,\ldots,N_{\text{FFT}}-1}$ and received signal Y_k can be expressed as $Y_k = H_k X_k + Z_k$, where Z_k is the frequency-domain noise sampled at the kth sub-carrier frequency, and H_k is the channel transfer-function sampled at the kth sub-carrier frequency. To estimate the channel, pilot symbols are used. Let us assume that every pth sub-carrier contains known pilot symbols $X_{pk}|_{kp=0,p,2p,\ldots,N_p-1}$, where N_p is the total number of pilot sub-carriers across frequency. Using the known pilot symbols X_{pk} and the received symbols Y_{pk} sampled at pilot sub-carriers, the channel transfer function $\hat{H}_k$ sampled at pilot sub-carriers can be written as follows:

$$\hat{H}_{pk} = \frac{Y_{pk}}{X_{pk}} + \frac{Z_{pk}}{X_{pk}} = H_{pk} + W_{pk} \tag{9.31}$$

where W_{pk} is a scaled-noise component at the pkth sub-carrier. Different methods such as one- or two-dimensional linear interpolation can be used to estimate the channel samples at other sub-carrier frequencies.

Channel estimation in OFDM systems is a two-dimensional problem, i.e., the channel transfer function or channel impulse response is a function of time and frequency variables. Due to the computational complexity of two-dimensional estimators, one-dimensional channel estimation methods are more practical. The idea behind one-dimensional estimators is to estimate the channel in one

Samsung Ex. 1005, Page 517 of 1147

dimension (e.g., frequency), and then estimate the channel in another dimension (e.g., time), thus obtaining a two-dimensional channel estimate. Linear interpolation is a simple method to estimate the channel from channel samples estimated at pilot sub-carriers. This is done by linearly interpolating the channel samples at the two nearest pilot sub-carriers. Although linear interpolation provides some limited noise reduction of the channel estimates at data locations (due to averaging function), it is the simplicity of the solution that is more attractive. It must be noted that averaging window length (i.e., the number of pilots contained in the averaging window) is inversely proportional to the coherence bandwidth of the channel. The one-dimensional linear interpolation method estimates the channel by interpolating the channel transfer function between $\hat{H}_{m,k}$ and $\hat{H}_{m,l}$ (interpolation across frequency), or between $\hat{H}_{m,k}$ and $\hat{H}_{n,k}$ (interpolation across time). The algorithm shown in Figure 9.47 illustrates the procedure for channel estimation across the time–frequency grid based on the known values of the channel at the reference signal locations. Depending on the relative location of the sub-carrier in the time–frequency grid, the channel is estimated via one-dimensional linear interpolation of channel values at neighboring pilots or merely copying or extrapolation of values at the neighboring pilot, if the previous or the next pilot does not exist [160].

The channel transfer function $H_{m,k}$ at time–frequency index (m, k) can be modeled as a linear weighed sum of two-dimensional basis functions evaluated at the kth sub-carrier (frequency index) and at the mth OFDM symbol (time index):

$$H_{m,k} = \sum_{n=0}^{N-1} \alpha_n \phi_n(m, k) \tag{9.32}$$

where $\phi_n(m, k)$ denotes the nth basis function sampled at the mth OFDM symbol and at the kth sub-carrier, α_n is the coefficient of the nth basis function, and N is number of basis functions used in the linear model. By taking a one-dimensional approach, the above two-dimensional problem is reduced to one-dimensional, if one of the time or frequency indices is kept unchanged:

$$H_k = \sum_{n=0}^{N-1} \alpha_n \phi_n(k) \tag{9.33}$$

The channel samples at pilot sub-carriers shown in Eq. (9.33) can be written as:

$$\hat{H}_{pk} = \sum_{n=0}^{N-1} \alpha_n \phi_n(pk) + W_{pk} \tag{9.34}$$

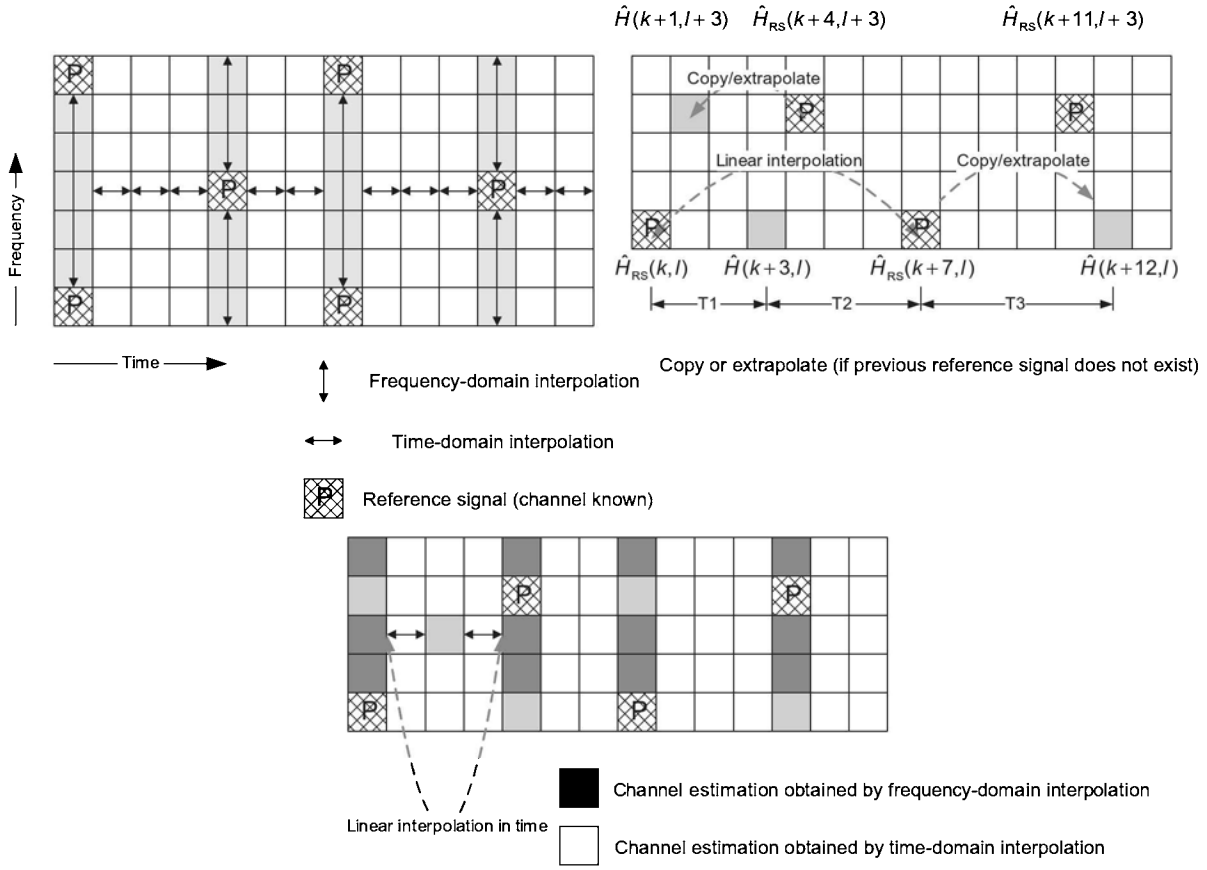


FIGURE 9.47

Two-dimensional channel estimation across frequency and time [176].

The above equation can be represented in matrix form as follows:

$$\begin{bmatrix} \hat{H}_0 \\ \hat{H}_p \\ \hat{H}_{2p} \\ \vdots \\ \hat{H}_{N_p-1} \end{bmatrix} = \begin{bmatrix} \phi_0(0) & \phi_1(0) & \phi_2(0) & \dots & \phi_{N-1}(0) \\ \phi_0(p) & \phi_1(p) & \phi_2(p) & \dots & \phi_{N-1}(p) \\ \phi_0(2p) & \phi_1(2p) & \phi_2(2p) & \dots & \phi_{N-1}(2p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_0(N_p-1) & \phi_1(N_p-1) & \phi_2(N_{p-1}) & \dots & \phi_{N-1}(N_{p-1}) \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{N-1} \end{bmatrix} + \begin{bmatrix} Z_0 \\ Z_p \\ Z_{2p} \\ \vdots \\ Z_{N_p-1} \end{bmatrix} \tag{9.35}$$

Alternatively, by using matrix notations, the above equation can be rewritten as follows:

$$\hat{\mathbf{H}} = \mathbf{\Phi}\boldsymbol{\alpha} + \mathbf{Z} \tag{9.36}$$

The least square (LS) estimate of the coefficients is calculated by minimizing the squared distance between the actual channel vector $\mathbf{H}$ and the estimated channel vector $\hat{\mathbf{H}}$ as follows:

$$\hat{\boldsymbol{\alpha}}_{\text{LS}} = (\mathbf{\Phi}^{\mathbf{H}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathbf{H}}\hat{\mathbf{H}} = \mathbf{\Phi}^{-1}\hat{\mathbf{H}} \tag{9.37}$$

where the superscript H denotes conjugate transpose operation, hence:

$$\hat{H}_k = \sum_{n=0}^{N-1} \hat{\alpha}_n \phi_n(k) \tag{9.38}$$

A number of different options exist while selecting the basis functions such as orthogonal polynomials, Fourier series, discrete cosine/sine transform function, etc. There are a number of reasons for using orthogonal polynomials including ease of calculation of $\mathbf{\Phi}^{\mathbf{H}}\mathbf{\Phi}$ since it becomes a diagonal matrix, if matrix $\mathbf{\Phi}$ is orthogonal, resulting in ease of calculation of $\hat{\alpha}_n's$ and the inverse matrix $(\mathbf{\Phi}^{\mathbf{H}}\mathbf{\Phi})^{-1}$ with minimal computational complexity. The degree of orthogonal polynomials can be increased without changes in recursive calculation of $\hat{\alpha}_n's$. When applying the polynomial-based channel estimation, a sliding window is typically used to estimate the channel. This is done to avoid use of high-order polynomials to estimate highly frequency-selective channels. The sliding window approach allows for local approximations to the channel transfer function using a window length larger than the polynomial order, providing a better estimate.

In a special case where the OFDM channel estimation symbols are transmitted periodically and all sub-carriers over an OFDM symbol are used as pilots, the channel estimation algorithm is more simplified. In this case, the receiver uses the estimated channel transfer function to decode the received data within the block until the next pilot symbol arrives. If $Y_k = H_k X_k + Z_k|_{k=0,1,2,\dots,N_{\text{FFT}}-1}$ denotes

the received signal at kth sub-carrier (pilot sub-carrier), the LS estimate of the channel transfer function in this case is given as:

$$\hat{\mathbf{H}}_{\mathbf{LS}} = (Y_1/X_1, Y_2/X_2, \ldots, Y_{N_{\text{FFT}}-1}/X_{N_{\text{FFT}}-1})^H = (Y_1, Y_2, \ldots, Y_{N_{\text{FFT}}-1})^H; X_k = 1, \forall k \tag{9.39}$$

The channel transfer function at non-pilot sub-carriers can be estimated using interpolation in time. For convenience, we assume that the pilot sub-carriers have unity magnitude.

An alternative approach to pilot-based channel estimation is the minimum mean squared error (MMSE) method. The MMSE channel estimator exploits the second-order statistics of the channel transfer function to minimize the mean-squared error. Let $\mathbf{R_{HH}} = E(\mathbf{HH}^{\mathrm{H}})$ and σ_Z^2 denote the autocorrelation matrix of the channel vector $\mathbf{H}$ and the additive white Gaussian noise variance, respectively. The MMSE estimate of the channel transfer function is given as:

$$\hat{\mathbf{H}}_{\text{MMSE}} = \mathbf{R_{HH}}[\mathbf{R_{HH}} + \sigma_Z^2(\mathbf{XX^H})^{-1}]^{-1}\hat{\mathbf{H}}_{\mathbf{LS}} = \mathbf{R_{HH}}[\mathbf{R_{HH}} + \sigma_Z^2\mathbf{I}]^{-1}\hat{\mathbf{H}}_{\mathbf{LS}}; \quad \mathbf{XX^H} = \mathbf{I} \tag{9.40}$$

The MMSE estimator performs much better than LS estimators, especially under the low SNR conditions. A major drawback of the MMSE estimator is its high computational complexity, particularly if matrix inversion is necessary every time the input data changes. Figure 9.48 shows an example where the

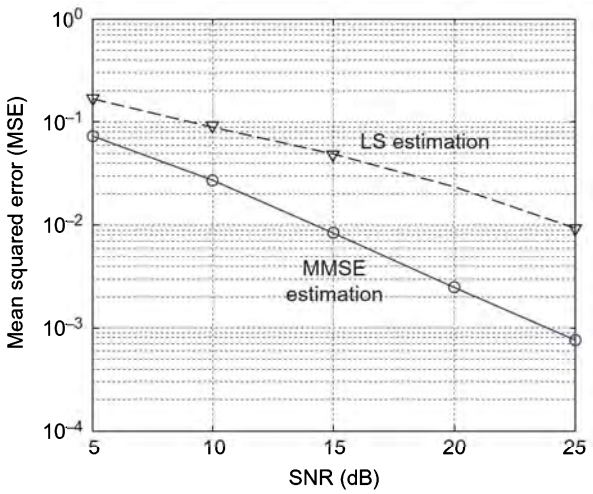


FIGURE 9.48

Example comparison of the LS and MMSE channel estimation algorithms [141].

performance of the LS and MMSE channel estimation algorithms has been compared. In this example $N_{FFT} = 64$ and the channel is modeled as a two-tap delay line filter. The superiority of MMSE over LS estimation in low SNR conditions can be seen in Figure 9.48.

9.10.3 Cell-specific reference signals

The cell-specific reference signals are used during initial cell selection for channel estimation and coherent detection of downlink physical channels. The measurements pertaining to scheduling and handover functions are based on the received signal strength of the CRS transmitted from antenna port zero (and antenna port one). The CRS sequence is a complex pseudo-random sequence, which is a length-31 Gold sequence.[11] The initialization seed for the second M-sequence used for Gold sequence generation is derived from the OFDM symbol number, slot number, and the physical cell identity. The seed is reinitialized at the beginning of each OFDM symbol within a slot. Pseudo-random sequences which are cell-specific are used to differentiate the reference signals transmitted from different cells. There is also a frequency shift that is applied to mapping of the CRS to resource elements to prevent the CRS from adjacent cells overlapping over the time–frequency grid. The generated sequence is transmitted in every slot and is spread throughout the channel bandwidth. The frequency of CRS transmission per slot and actual mapping to resource elements depend on the antenna configuration (Figure 9.49). The CRS symbols between antennas from the same cell are time-division multiplexed; the resource elements used for CRS symbols on one antenna are reserved and are not used on other antenna ports.

The CRSs are transmitted in all downlink subframes in a cell supporting PDSCH transmission. The CRSs are transmitted on one or several of antenna ports 0–3, and are only defined for sub-carrier spacing of 15 kHz. The reference-signal sequence $r_l^{n_s}(m)$ is defined as $r_l^{n_s}(m) = (1/\sqrt{2})(1 - 2c(2m)) + j(1/\sqrt{2})(1 - 2c(2m+1))$, $\forall m = 0, 1, \ldots, 2N_{RB}^{DL} - 1$, where index n_s denotes the slot number within a radio frame, $l = 0, 1, \ldots, 6$ (or 5 in case of an extended cyclic prefix) represents the OFDM symbol number within slot n_s, and $c(i)$ is a pseudo-random sequence, which is initialized with $c_{init} = 2^{10}[7(n_s + 1) + l + 1](2N_{ID}^{cell} + 1) + 2N_{ID}^{cell} + N_{CP}$ at the start

[11] A Gold sequence is a type of binary sequence that is often used in telecommunication and satellite navigation. Gold sequences have bounded small cross-correlations within a set. A set of Gold sequences consists of $2^n - 1$ sequences each with a period of $2^n - 1$. A set of Gold sequences can be generated by taking two maximum-length sequences of the same length $2^n - 1$ such that their absolute cross-correlation is less than or equal to $2^{(n+2)/2}$, where n is the size of the linear feedback shift register used to generate the maximum length sequence [130]. The set of $2^n - 1$ (logical) exclusive OR of the two sequences in their various phases is a set of Gold codes. The highest absolute cross-correlation in this set of codes is $2^{(n+2)/2} + 1$ for even n, and $2^{(n+1)/2} + 1$ for odd n. The exclusive OR of two Gold codes from the same set is another Gold code with arbitrary phase [130].

Samsung Ex. 1005, Page 522 of 1147

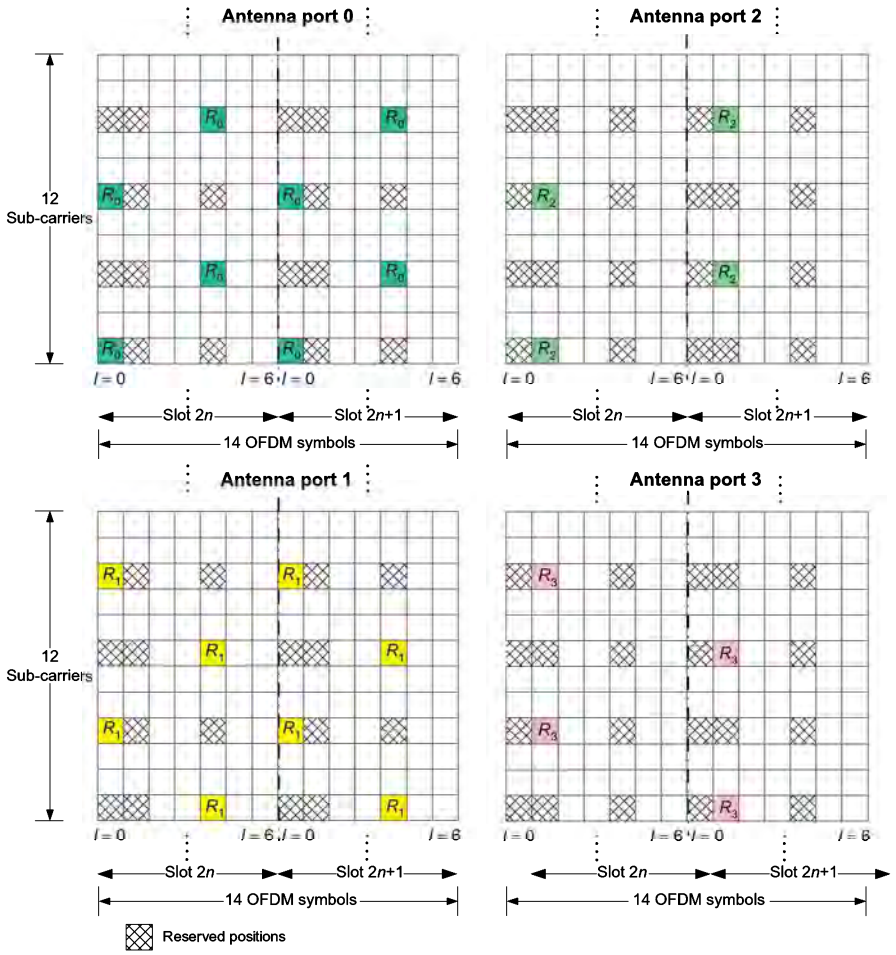


FIGURE 9.49

Mapping of cell-specific reference signals to antenna ports 0, 1, 2, and 3 for four transmit antennas over a resource block using a normal cyclic prefix [3].

of each OFDM symbol.[12] Note that $N_{CP} = 1$ for a normal cyclic prefix and $N_{CP} = 0$ for an extended cyclic prefix. The reference signal sequence $r_l^{n_s}(m)$ is mapped to complex-valued modulation symbols a_{kl}^p that are designated as reference symbols for

[12]Pseudo-random sequences are defined by a length-31 Gold sequence. The output sequence $c(n), \forall n = 0, 1, \ldots, M-1$ of length M is defined by $c(n) = (x_1(n + N_C) + x_2(n + N_C))\text{mod } 2$ in which $x_1(n + 31) = (x_1(n + 3) + x_1(n))\text{mod}2, x_2(n + 31) = (x_2(n + 3) + x_2(n + 2) + x_2(n + 1) + x_2(n))\text{mod } 2$, and $N_C = 1600$. The first m-sequence is initialized with $x_1(0) = 1, x_1(n) = 0, n = 1, 2, \ldots, 30$. The second M-sequence is initialized by $c_{\text{init}} = \sum_{i=0}^{30} x_2(i)2^i$ with the value depending on the application of the sequence [3].

antenna port p in slot n_s as $a_{kl}^{p} = r_{l}^{n_s}(m')$, where the parameters are given as follows [3]:

$$k = 6m + (v + v_{\text{shift}}) \bmod 6; \quad l = \begin{cases} 0, N_{s}^{\text{DL}} - 3 & \text{if } p \in \{0, 1\} \\ 1 & \text{if } p \in \{2, 3\} \end{cases}$$

$$m = 0, 1, \ldots, 2N_{\text{RB}}^{\text{DL}} - 1; \qquad m' = m + \max(N_{\text{RB}}^{\text{DL}}) - N_{\text{RB}}^{\text{DL}} \tag{9.41}$$

The variables v and v_{shift} define the position in the frequency-domain for different reference signals, where v is given by the following expression [3]:

$$v = \begin{cases} 0 & \text{if } p = 0 \text{ and } l = 0 \\ 3 & \text{if } p = 0 \text{ and } l \neq 0 \\ 3 & \text{if } p = 1 \text{ and } l = 0 \\ 0 & \text{if } p = 1 \text{ and } l \neq 0 \\ 3(n_s \bmod 2) & \text{if } p = 2 \\ 3 + 3(n_s \bmod 2) & \text{if } p = 3 \end{cases} \tag{9.42}$$

A cell-specific frequency shift is given by $v_{\text{shift}} = N_{\text{ID}}^{\text{cell}} \bmod 6$ (Figure 9.50). As shown in Figure 9.50, the resource element allocation for CRSs cycles once every six physical cell identity values, i.e., cell identity values that are integer multiples of 6 has the same resource element allocation as cell identity 0. This corresponds to one cycle for every two physical layer cell-identity groups. The resource elements

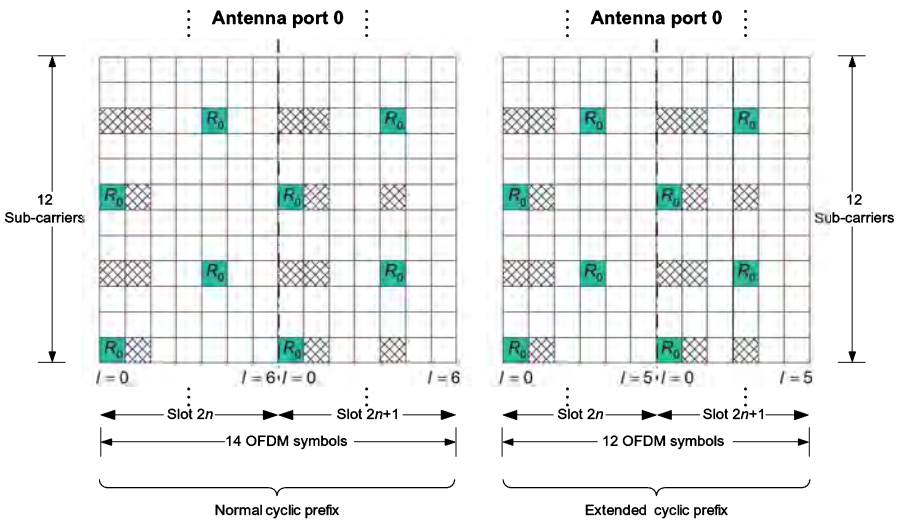


FIGURE 9.50

Cell-specific reference signal mapping for normal and extended cyclic prefix subframes [3].

(k, l) used for transmission of CRSs on an any of the antenna ports in a slot are not used for any transmission on another antenna port in the same slot and are set to zero. In an MBSFN subframe, CRSs are only transmitted in the non-MBSFN region of the MBSFN subframe. Figure 9.49 shows the CRS mapping to the resource grid. The procedure for generation of the CRSs is illustrated in Figure 9.51.

Given a different number of available OFDM symbols in a subframe with an extended cyclic prefix, the mapping of the CRSs is different than that for subframes with a normal cyclic prefix. Figure 9.50 illustrates and compares the CRS mapping in subframes with a normal and extended cyclic prefix associated with antenna port 0 and 4 transmit antennas. While the number of resource elements mapped with reference signals is the same, the CRS overhead in subframes with an extended cyclic prefix is higher due to a lower number of total sub-carriers in a resource block.

As shown in Figure 9.51, the CRSs are based on length-31 Gold pseudo-random sequences. The length-31 Gold pseudo-random sequence is generated with the seed c_{init}, based on the slot number, symbol number, cell identity, and cyclic prefix type. While the sequence itself is 230 bits long, the number of bits from the sequence selected for transmission is based on the largest channel bandwidth, which is currently 20 MHz. Therefore, $4 \max(N_{\text{RB}}^{\text{DL}})$ bits are selected from

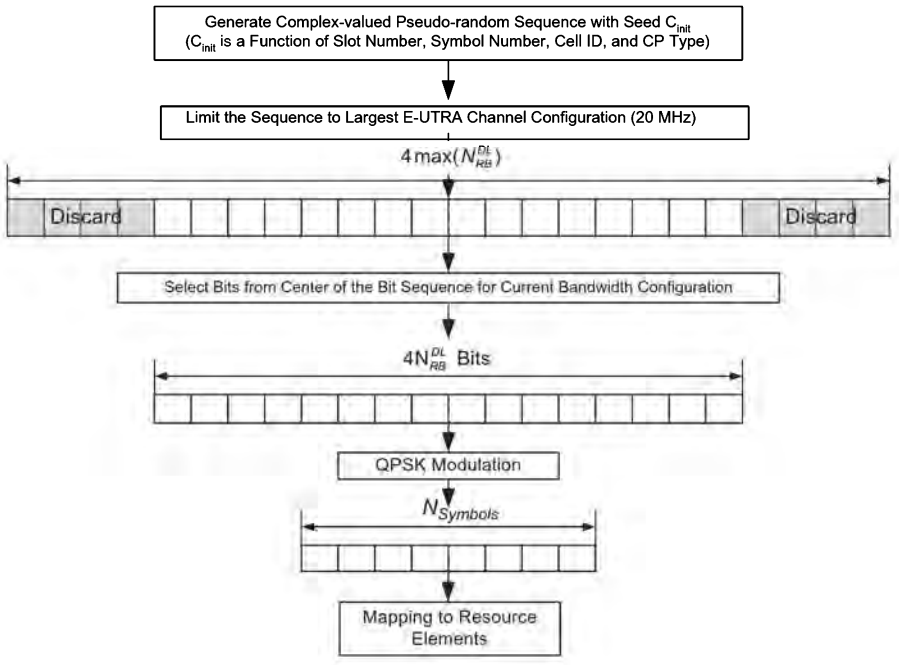


FIGURE 9.51

Generation of cell-specific reference signals.

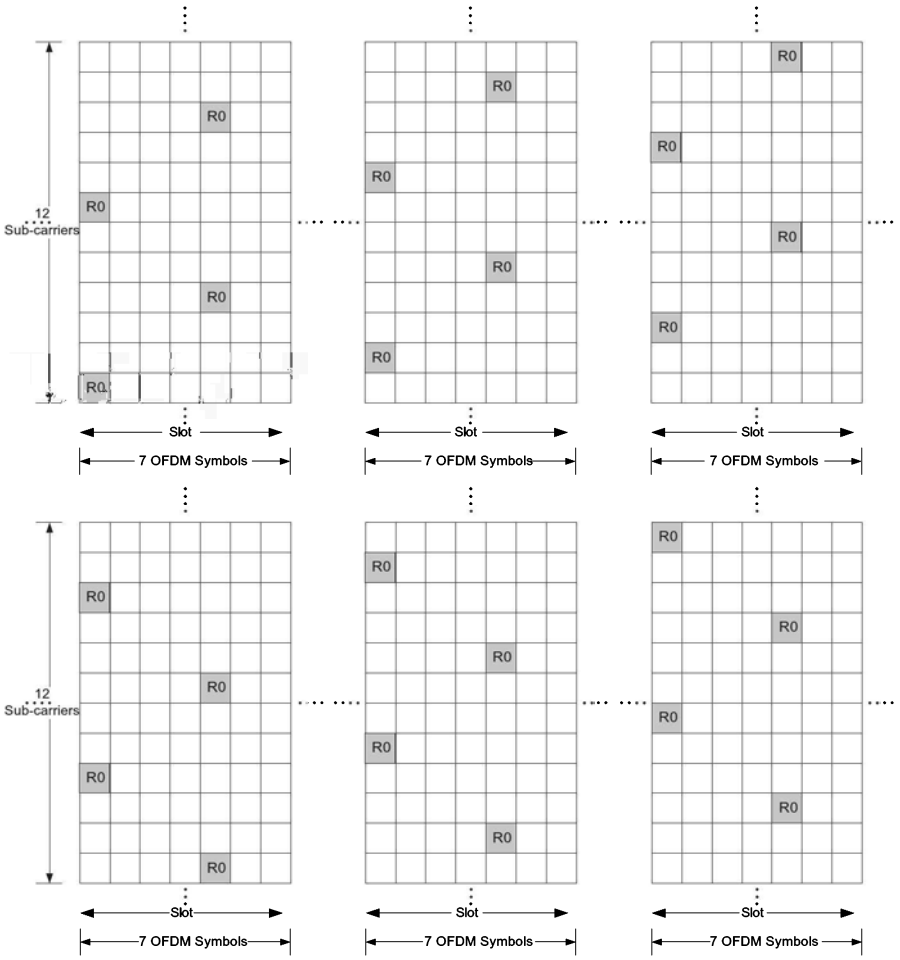


FIGURE 9.52

Dependency of cell-specific reference signals location on physical cell identifier [13].

the sequence where $\max(N_{\mathrm{RB}}^{\mathrm{DL}})$ is the maximum number of resource blocks in 20 MHz. From 4 $\max(N_{\mathrm{RB}}^{\mathrm{DL}})$, $4N_{\mathrm{RB}}^{\mathrm{DL}}$ bits are selected, where $N_{\mathrm{RB}}^{\mathrm{DL}}$ is the number of resource blocks corresponding to the system bandwidth. The bits are selected from the center of the truncated sequence for the maximum channel configuration. This ensures that the same reference symbols are present around PBCH regardless of channel bandwidth. The PBCH uses the six resource blocks (72 subcarriers) at the center of the channel. The UE can use these symbols for channel estimation for PBCH detection without knowing the overall channel bandwidth. The channel bandwidth in terms of resource blocks is determined from the

PBCH. After PBCH is decoded, the UE determines the length of the reference signal used for the current channel configuration.

9.10.3.1 Received signal strength indicator measurement

Received signal strength indicator (RSSI) is the total power that UE observes across the entire system bandwidth. This includes the main signal and co-channel non-serving cell signal, adjacent channel interference, and the thermal noise within the specified frequency band. This is the power of the received signal prior to demodulation, thus the UE can measure this power without any synchronization and demodulation [6].

9.10.3.2 Reference signal received power measurement

Reference signal received power (RSRP) is defined as the linear average over the power contributions of the resource elements that carry cell-specific reference signals within the considered measurement bandwidth. For the purpose of RSRP calculation, the R_0 cell-specific reference signals are used; however, if the UE can reliably detect R_1 cell-specific reference signals, then it may use R_1 in addition to R_0 in order to calculate RSRP. The reference point for the RSRP measurement is the antenna connector of the UE. If the UE uses a receive diversity scheme, the reported RSRP value must not be lower than individual RSRP measured at any of the diversity branches. There are requirements for both absolute and relative RSRP. The absolute requirements range from ± 6 to ± 11 dB, depending on the noise level and environmental conditions. Measuring the difference in RSRP between two cells on the same frequency (intra-frequency measurement) is a more accurate operation for which the requirements vary from ± 2 to ± 3 dB. The requirements widen again to ± 6 dB when the cells are on different frequencies (inter-frequency measurement) [6].

9.10.3.3 Reference signal received quality measurement

Reference signal received quality (RSRQ) is defined as the ratio $\text{RSRQ} = N_{\text{RB}}\,\text{RSRP}/\text{RSSI}_{E\text{-UTRA}}$ (dB), where N_{RB} is the number of resource blocks over which the measurement is conducted. Note that the value of RSRQ in decibels is always negative, since the denominator is larger than the numerator. The measurements in the numerator and denominator are conducted over the same set of resource blocks. The E-UTRA carrier RSSI comprises the linear average of the total received power observed only in OFDM symbols containing the cell-specific reference symbols for antenna port zero over N_{RB} resource blocks in the measurement bandwidth by the UE from all sources, including co-channel serving and non-serving cells, adjacent channel interference, and thermal noise. If higher-layer signaling indicates certain subframes for performing RSRQ measurements, then RSSI is measured over all OFDM symbols in the designated subframes. The reference point for the RSRQ is the antenna connector of the UE. If the UE uses a receive diversity scheme, the reported RSRP value must not be lower than individual RSRQ measured at any of the diversity branches.

Table 9.10 RSRQ Mapping [6]

Reported RSRQ Value	Measured Quantity (dB)
RSRQ_00	RSRQ < −19.5
RSRQ_01	−19.5 ≤ RSRQ < −19
RSRQ_02	−19 ≤ RSRQ < −18.5
...	...
RSRQ_32	−4 ≤ RSRQ < −3.5
RSRQ_33	−3.5 ≤ RSRQ < −3
RSRQ_34	−3 ≤ RSRQ

Measuring RSRQ becomes particularly important near the cell-edge when decisions need to be made regardless of absolute RSRP to perform a handover to the target cell. Reference signal receive quality is used only during connected states. Intra- and inter-frequency absolute RSRQ accuracy varies from ±2.5 to ±4 dB, which is similar to the inter-frequency relative RSRQ accuracy of ±3 to ±4 dB.

It must be noted that RSRQ measurement provides additional information when RSRP is not sufficient to make a reliable handover or cell reselection decision. The RSRQ is the ratio of the RSRP and the RSSI, which depends on the measurement bandwidth, and thereby the number of resource blocks. The RSSI value includes the total received wideband power from all interference and thermal noise sources. Since RSRQ combines signal strength as well as interference level, its value further assists mobility management in E-UTRA. The UE typically measures RSRP or RSRQ per instructions from the network (via RRC messages) and reports the values back to the eNB. When it reports the RSRQ value, it does not use the real RSRQ value; rather it sends a non-negative value ranging from 0 to 34, and each of these values are mapped to a specific range of real RSRQ values as given in Table 9.10. The reporting range of RSRQ is defined from −19.5 dB to −3 dB with 0.5 dB resolution. As an example, assume that only reference signals are transmitted in a resource block and that data, noise, and interference are not considered. In this case, the RSRQ is equal to −3 dB. If cell-specific reference signals and the sub-carriers carrying data are equally powered, the RSRQ is equal to 1/12 or −10.79 dB.

9.10.4 UE-Specific reference signals associated with PDSCH

It was mentioned that cell-specific reference signals are transmitted on each antenna port. An antenna port is generally used as a generic term for signal transmission under identical channel conditions. The symbols that are transmitted via identical antenna ports are subject to the same channel conditions. In order to determine the characteristics of the channel from an antenna port, a UE must carry out a separate channel estimation for each antenna port. Therefore, separate

reference signals that are suitable for estimating the respective channel are defined for each antenna port. There is one resource grid per antenna port. The cell-specific reference signals are used for both coherent detection and mobility-related measurements. Their structure in time and frequency ensures sufficient channel estimation accuracy. An LTE-Advanced eNB is required to support LTE Rel-8 user terminals. Since the structure of PCFICH, PHICH, PDCCH, PBCH, and PDSCH (transmit diversity only) is unchanged in Rel-10, the support of CRS is also required for LTE-Advanced UEs in order to detect these essential physical channels. The UE-specific reference signals are supported for transmission of PDSCH, and are transmitted on antenna ports $p = 5$, $p = 7$, $p = 8$, or $p = 7, 8, \ldots, N_l + 6$, where N_l is the number of spatial layers used for transmission of PDSCH. The UE-specific reference signals or DM-RS can be precoded and support non-codebook-based precoding. The DM-RS pattern for a higher number of layers has been extended from dual-layer format of transmission mode 8 in Rel-9 to eight layers in Rel-10. The CSI-RS is transmitted on each physical or virtual antenna port, and is only used for channel state measurement/reporting purposes. Its channel estimation accuracy can be relatively poorer compared to that of DM-RS. In Rel-8, the CRS is used for both demodulation and measurement purposes. When using precoding in conjunction with CRSs, the precoding information must be sent to the UE, and the CRS power is boosted to support various types of geometry. The Rel-8 E-UTRA downlink MIMO transmission modes are based on CRS. Figure 9.53 shows the CRS-based procedure for precoding where the cell-specific reference signals are inserted after the data is precoded. Therefore, the cell-specific reference signals do not contain information about the

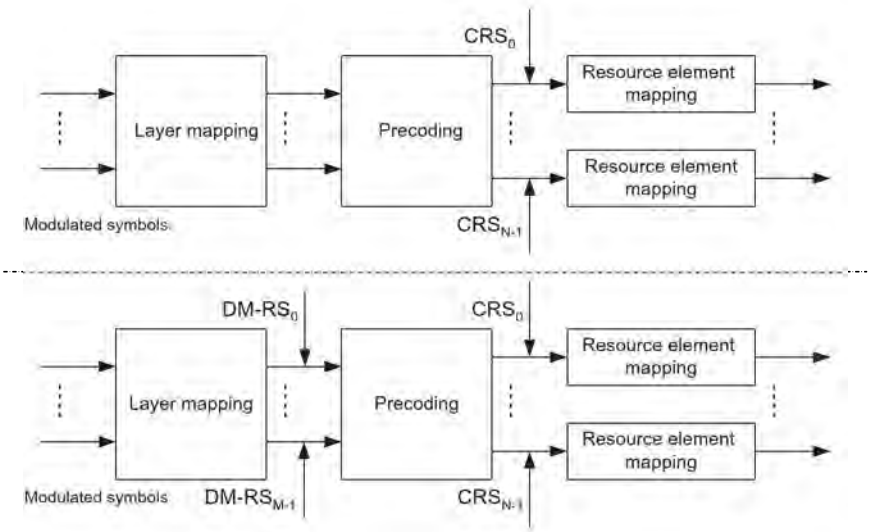


FIGURE 9.53

Comparison of the CRS and DM-RS usage in Rel-8 and Rel-10 of E-UTRA [110].

precoding scheme, and codebook information must be separately sent to the UE. In contrast, in LTE-Advanced, downlink MIMO transmission is based on DM-RS and CSI-RS. Figure 9.53 shows the DM-RS-based procedure for precoding where UE-specific reference signals are precoded in the same way as the data. Therefore, no separate precoding information is sent to the UE. Note that in LTE Rel-10, DM-RS is used for demodulation and CSI-RS is used for channel state measurement purposes. Unlike CRS-based precoding, the precoding information is not needed for demodulation.

It was stated earlier that UE-specific reference signals are used for transmission of the downlink shared channel and are transmitted on antenna port(s) $p = 5$, $p = 7$, $p = 8$ or $p = 7, 8, \ldots, N_l + 6$, where N_l is the number of layers (up to eight). It must be noted that the UE-specific reference signals on any antenna port are a valid reference for data demodulation, if the data transmission is associated with the corresponding antenna port. The UE-specific reference signals are exclusively transmitted on the resource blocks to which the corresponding user data is mapped. The UE-specific reference signal is not mapped to resource elements (k, l) that are reserved for transmission of control signaling, physical channels, or physical signals.

For antenna port 5, the UE-specific reference-signal sequence $r_{n_s}(m)$ is defined by $r_{n_s}(m) = (1/\sqrt{2})[1 - 2c(2m)] + j(1/\sqrt{2})[1 - 2c(2m + 1)]$, $\forall m = 0, 1, \ldots, 12N_{RB} - 1$, where N_{RB} denotes the number of resource blocks allocated to the corresponding data transmission. The pseudo-random sequence $c(i)$ was defined earlier. The pseudo-random sequence generator is initialized with $c_{init} = (\lfloor n_s/2 \rfloor + 1)(2N_{ID}^{cell} + 1)2^{16} + n_{RNTI}$ at the start of each subframe where n_{RNTI} is the 16-bit radio network temporary identifier (RNTI) of the UE. For antenna ports $p \in \{7, 8, \ldots, N_l + 6\}$, the UE-specific reference-signal sequence $r_{n_s}(m)$ is defined by $r_{n_s}(m) = (1/\sqrt{2})[1 - 2c(2m)] + j(1/\sqrt{2})[1 - 2c(2m + 1)]$, where $m = 0, 1, \ldots, 12N_{RB}^{DL} - 1$ (for normal cyclic prefix) or $m = 0, 1, \ldots, 16N_{RB}^{DL} - 1$ (for extended cyclic prefix). The pseudo-random sequence $c(i)$ is defined similar to the previous section. Prior to LTE Rel-11, the pseudo-random sequence generator was initialized with $c_{init} = (\lfloor n_s/2 \rfloor + 1)(2N_{ID}^{cell} + 1)2^{16} + n_{SCID}$ at the start of each subframe. The value of the scrambling identity n_{SCID} is zero unless specified otherwise. For a data transmission on ports 7 or 8, n_{SCID} is indicated by the DCI format 2B or 2C. In the case of DCI format 2B, $n_{SCID} = 0$ if the scrambling identity field is zero. It must be noted that CQI estimation at the terminal is always based on cell-specific reference signals. In LTE Rel-11, and as a result of the introduction of the new PDCCH structure based on the UE-specific reference signals, the initialization of the DM-RS sequences associated with antenna ports $p \in \{7, 8, \ldots, N_l + 6\}$ was modified as follows: $c_{init} = (\lfloor n_s/2 \rfloor + 1)\left(2N_{ID}^{(n_{SCID})} + 1\right)2^{16} + n_{SCID}$, where $N_{ID}^{(i)}|_{i=0,1}$ is defined as $N_{ID}^{(i)} = N_{ID}^{cell}$, if no value for $N_{ID}^{DM\text{-}RS_i}$ is conveyed by higher layers, or if DCI format 1A is used for the DCI associated with PDSCH transmission; otherwise $N_{ID}^{(i)} = N_{ID}^{DM\text{-}RS_i}$ [3].

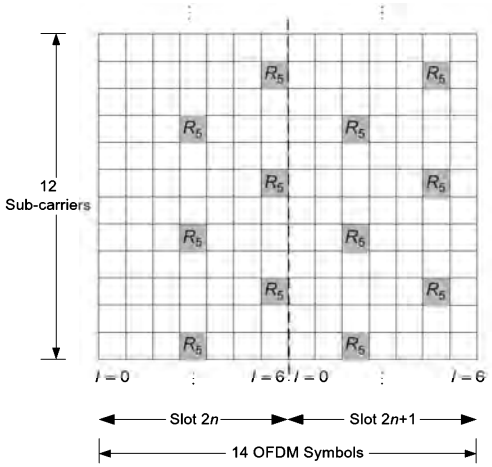


FIGURE 9.54

Mapping of UE-specific reference signals to antenna port 5 [3].

In a PRB with index n_{PRB} allocated for data transmission on antenna port 5, and assuming the normal cyclic prefix, the DM-RS $r_{n_s}(m)$ is mapped to the complex-valued modulation symbols a_{kl}^{p} over slot n_s as $a_{kl}^{5} = r_{n_s}(3l'N_{\text{RB}} + m')$, where $m' = 0, 1, \ldots, 3N_{\text{RB}} - 1$ denotes the number of UE-specific reference signal resource elements within a respective OFDM symbol allocated to user data transmission, $l' = \{0, 1\}$ for even-numbered slots, and $l' = \{2, 3\}$ for odd-numbered slots.

The UE-specific reference signal pattern for up to two layers, or alternatively rank-2 transmission, is identical to that of LTE Rel-9 to maintain backward compatibility. For up to four layers, the DM-RS pattern is obtained by extending LTE Rel-9 UE-specific reference signal pattern using a combination code-division and frequency-division multiplexing, as shown in Figure 9.55. The four precoded layers (antenna ports 7–10) are grouped into two groups of two resource elements that are spread with the same length-2 Walsh–Hadamard[13] orthogonal cover codes (OCCs) as in Rel-9. The UE-specific reference signals in different groups are frequency-multiplexed on adjacent sub-carriers. This pattern has been shown to provide reasonable performance for a variety of channel conditions and UE speeds [121].

[13]In coding theory, the Walsh–Hadamard code is a linear code over a binary alphabet that maps messages of length 2^n. The Walsh–Hadamard code is unique in that each non-zero codeword has a Hamming weight of exactly 2^{n-1}, which implies that the distance of the code is also 2^{n-1} resulting in a $[2^n, n, 2^{n-1}]_2$ code. The Walsh–Hadamard code is locally decodable, which enables recovery of parts of the original message with high probability, while only looking at a small fraction of the received codeword [27,40].

Samsung Ex. 1005, Page 531 of 1147

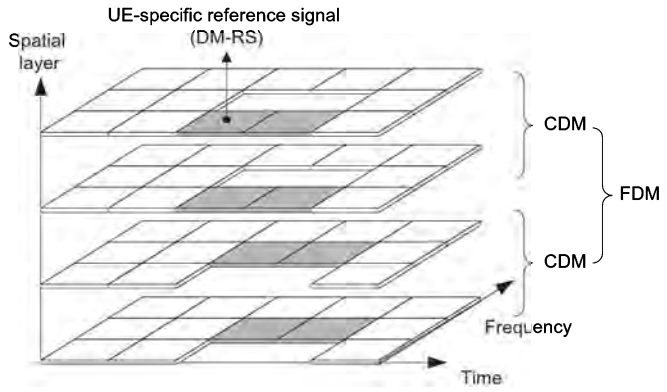


FIGURE 9.55

Code and frequency division multiplexing of the UE-specific reference signals [177].

For rank-8 transmission on antenna ports 7–14, the UE-specific reference signal pattern of Figure 9.55 is further extended using a hybrid CDM/FDM approach with two CDM groups each using a Walsh–Hadamard OCC of length 4, as shown in Figure 9.56. The UE-specific reference signals for each layer are spread across four resource elements, all having the same frequency location but different time locations within the subframe. This approach maintains the balanced power property of the rank-4 UE-specific pilot design and is optimized for pedestrian mobility class, which is the most likely use case for eight-layer transmission. The OCCs of length two are embedded in the OCC of length four to ensure backward compatibility with LTE Rel-9. The mapping of the OCCs to resource elements is shown in Figure 9.56. The mapping for odd and even resource blocks (in the frequency-domain) is different in order to provide orthogonality in both time and frequency dimensions. Such orthogonality improves the performance of channel estimation in high Doppler scenarios, and reduces the inter-carrier interference across the two CDM groups in the presence of frequency offsets. The OCC mapping also enables power balancing across OFDM symbols by peak power randomization [14]. The reference sequence for up to eight layers is the same as in Rel-9, except for antenna ports $AP_p|_{p=9,10,\ldots,14}$ where only one sequence initialization is possible.

In mathematical terms, for antenna ports $AP_p|_{p=7,8,\ldots,N_l+6}$, in a PRB with frequency-domain index n_{PRB} assigned for the corresponding user data transmission, part of the reference sequence $r(m)$ is mapped to complex-valued modulation symbols a_{kl}^p in a subframe (assuming normal cyclic prefix) as follows: $a_{kl}^p = w_p(l')\, r[3l'\max(N_{\text{RB}}^{\text{DL}}) + 3n_{\text{PRB}} + m']$, where $w_p(i) = \overline{w}_p(i)$ if $(m' + n_{\text{PRB}}) \bmod 2 = 0$ or $w_p(i) = \overline{w}_p(3-i)$ if $(m' + n_{\text{PRB}}) \bmod 2 = 1$, where $\overline{w}_p(i)$ is given in Table 9.11. The UE-specific reference signal patterns for different transmission ranks (number of spatial layers) are shown in Figure 9.57.

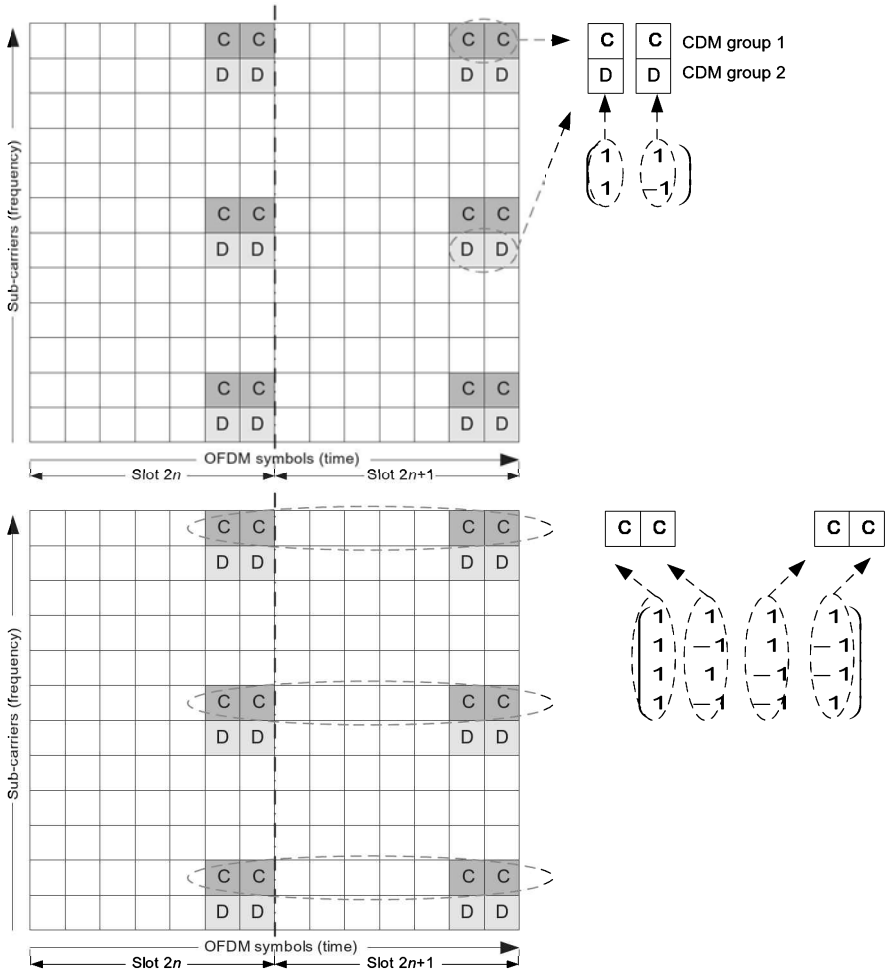


FIGURE 9.56

Length-2 and length-4 OCC mapping and DM-RS structure [121].

9.10.5 CSI reference signals

The new multi-user MIMO techniques that were proposed as part of Rel-10 and beyond require some level of CSI knowledge at the base station so that the system can dynamically adapt to the radio channel conditions experienced by the user terminals in order to optimize the performance. In TDD systems, this information is intuitively obtained from the uplink direction based on the reciprocity of the wireless channel, provided that the channel fading is sufficiently slow and transceivers at the base and mobile stations are calibrated, due to the fact that the

Table 9.11 Orthogonal Cover Code of Length 4 [3]

Antenna Port p	$[\bar{w}_p(0)$	$\bar{w}_p(1)$	$\bar{w}_p(2)$	$\bar{w}_p(3)]$
7	[+1	+1	+1	+1]
8	[+1	−1	+1	−1]
9	[+1	+1	+1	+1]
10	[+1	−1	+1	−1]
11	[+1	+1	−1	−1]
12	[−1	−1	+1	+1]
13	[+1	−1	−1	+1]
14	[−1	+1	+1	−1]

same carrier frequency is used for transmission and reception. On the other hand, due to the asymmetry of FDD systems, feedback information over the uplink is required. Full CSI could cause an excessive overhead in some scenarios, thus quantized or statistical CSI are used in practice. In addition, terminal mobility can cause further degradation of the system performance, if the CSI arriving at the eNB is already outdated. Multi-antenna techniques in a multi-user scenario have the role of delivering streams of data in a spatially multiplexed manner to different users in such a way that all the degrees of freedom of a MIMO system are utilized. The idea is to perform intelligent space-division multiple access or multi-user beamforming so that the transmission of the base station is adapted to each user's instantaneous channel conditions in order to achieve the maximum gain in the direction of that user. The base station collects information on the CSI of each UE and subsequently decides on the optimal resource allocation strategy for each user.

Channel state information reference signals were introduced as part of Rel-10 to enable channel state measurement and reporting by UEs for up to eight downlink transmit-antennas to facilitate eNB precoding functions in transmission mode 9. LTE Rel-10 supports transmission of CSI-RS for 1, 2, 4, and 8 transmit antennas. The CSI-RSs also enable the UE to estimate the CSI for multiple neighbor cells in addition to its serving cell, to support multi-cell cooperative transmission schemes. The following principles have been incorporated in the design of the CSI-RS [14]:

- It is desirable to place CSI-RS at locations with uniform spacing in the frequency-domain.
- It is desirable to minimize the number of subframes containing CSI-RS in the time-domain so that a UE can estimate the CSI for different antenna ports and different cells with a minimal wake-up (active state) duty cycle when the UE is in DRX mode to reduce UE power consumption.
- The overall CSI-RS overhead involves a trade-off between accurate CSI estimation for efficient operation and minimizing the impact on legacy UEs

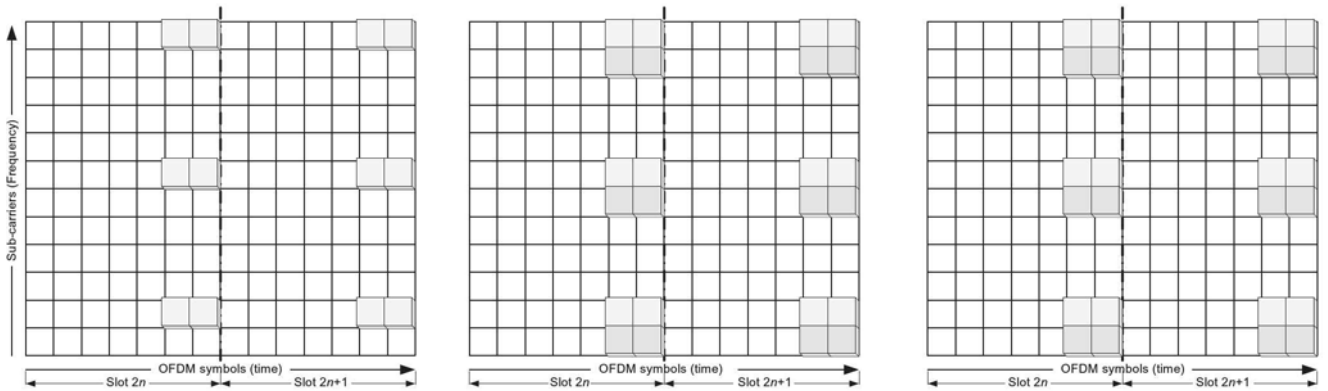


FIGURE 9.57

Mapping of UE-specific reference signals to resource blocks for different numbers of layers [177].

which are unaware of the presence of CSI-RS. Furthermore, the data sub-carriers of the legacy UE are punctured by the CSI-RS transmissions. Figure 9.58 shows that a CSI-RS density of one resource element per resource block per antenna port is a reasonable choice, as the throughput degradation compared to ideal CSI estimation is negligible.

- The CSI-RSs of different antenna ports within a cell, and ideally from other cells, should be orthogonally multiplexed to enable accurate CSI estimation.
- To ensure backward compatibility, CSI-RSs should not be mapped to the resource elements used for CRSs and control channels, as well as avoiding resource elements used for the Rel-10 UE-specific reference signals.

A set of CSI-RSs were designed in Rel-10, taking the above design considerations into account. An example CSI-RS pattern for 1, 2, 4, and 8 antenna ports is shown in Figure 9.58. The OCC codes of length 2 are used, so that CSI-RSs on two antenna ports share two resource elements on a given sub-carrier. The pattern shown in Figure 9.58 is applicable to frame structure type 1 and frame structure type 2 [121].

The CSI-RS patterns follow a nested structure, meaning that the resource elements used in the case of two CSI-RS antenna ports are a subset of those used for four and eight antenna ports; this helps to simplify the implementation. The total number of supported CSI-RS patterns is 40, which can be used to give a frequency-reuse factor of 5 between cells with eight antenna ports per cell, or a factor of 20 in the case of two antenna ports. It can be seen that collisions may occur with resource elements used for the DM-RS antenna ports defined in Rel-8 for PDSCH transmission mode 7. Therefore, it may be desirable to avoid scheduling UEs in transmission mode 7 in subframes containing CSI-RS. This pattern is designed to avoid collisions with the Rel-8 UE-specific reference signal antenna port, as this port is more suited to TDD operation where channel reciprocity can be more effectively exploited to support beamforming. However, this pattern only offers frequency-reuse factors between 3 and 12 depending on the number of antenna ports per cell [14].

The CSI-RS configuration is UE-specific. When the CSI-RS is configured, it is only present in certain subframes conforming to a given duty cycle and subframe offset. The duty cycle and offset of the subframes containing CSI-RSs and the CSI-RS pattern used in those subframes are provided to a Rel-10 UE through RRC signaling. The duty cycle and subframe offset are jointly coded, while the CSI-RS pattern is configured independently of these parameters. It should be noted that in subframes containing CSI-RSs, the rate-matching for PDSCH transmissions to Rel-10 UEs is considerate of the resource elements used for CSI-RS and the encoded PDSCH data is only mapped to the neighboring sub-carriers, whereas PDSCH transmissions to the legacy UEs are punctured by CSI-RS transmissions. Therefore, the density and periodicity of the CSI-RSs should be chosen

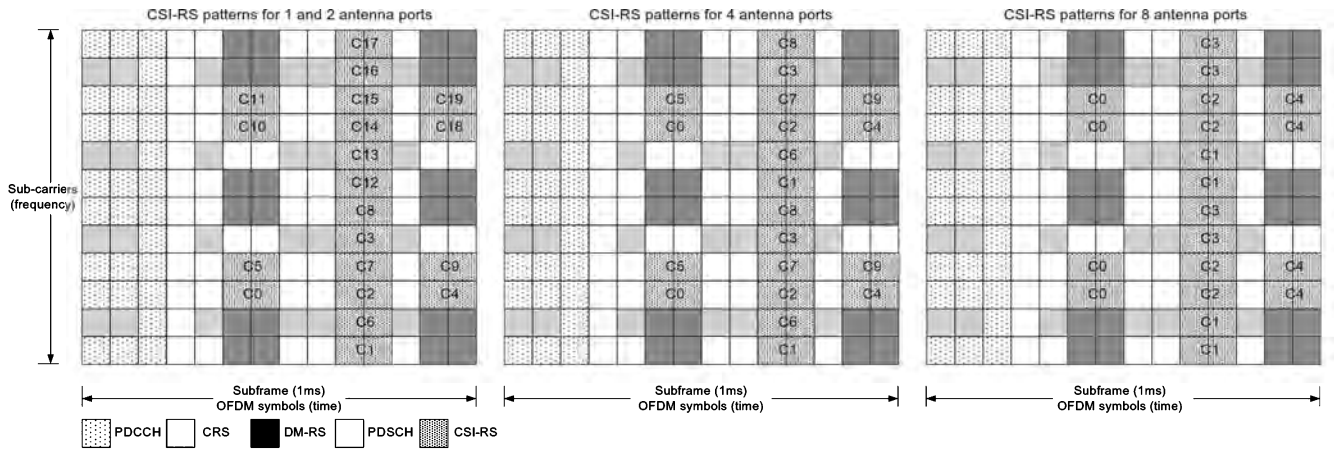


FIGURE 9.58

CSI-RS patterns for 1, 2, 4, and 8 transmit antennas [3,121].

such that the impact of puncturing[14] on the performance of these users is negligible.

In the context of cooperative MIMO, it may be possible to improve the performance of channel estimation, and especially interference estimation, by coordinating CSI-RS transmissions across multiple cells. In Rel-10 it is possible to mute a specific set of sub-carriers in data transmissions from a cell. The locations of these resource elements can be chosen to avoid colliding with CSI-RS transmissions from other cells, thus improving the inter-cell interference measurement quality. The muting pattern is signaled to the UEs using a 16-bit bitmap, where each bit corresponds to a 4-port CSI-RS configuration for frame structure type 1 or type 2. As a result, all the resource elements in the selected 4-port CSI-RS configuration are muted and the UE can assume zero transmission power on those resource elements, unless they are used to transmit CSI-RSs. In subframes in which CSI-RS muting is configured, data transmissions for Rel-10 UEs are rate-matched around the muted resource elements in the same way as for CSI-RSs. Data transmitted to Rel-8/9 UEs is punctured at the muted resource elements [15].

The CSI-RSs are transmitted on one, two, four, or eight antenna ports $AP_p = \{p = 15\}, \{p = 15, 16\}, \{p = 15, \ldots, 18\}, \{p = 15, \ldots, 22\}$ and are only defined for sub-carrier spacing of 15 kHz. The reference signal sequence is defined as $r_l^{n_s}(m) = (1/\sqrt{2})[1 - 2c(2m)] + \mathrm{j}(1/\sqrt{2})[1 - 2c(2m+1)]$, $\forall m = 0, 1, \ldots, \max(N_{\mathrm{RB}}^{\mathrm{DL}}) - 1$, where n_s is the slot number within a radio frame and $l = 0, 1, \ldots, 6$ (or 5 for an extended cyclic prefix) is the OFDM symbol number within the slot. The pseudo-random sequence $c(i)$ was defined in the previous sections. The pseudo-random sequence generator is initialized with $c_{\mathrm{init}} = 2^{10}(7(n_s + 1) + l + 1)(2N_{\mathrm{ID}}^{\mathrm{CSI}} + 1) + 2N_{\mathrm{ID}}^{\mathrm{CSI}} + N_{\mathrm{CP}}$ at the start of each OFDM symbol where $N_{\mathrm{CP}} = 1$ for a normal cyclic prefix and $N_{\mathrm{CP}} = 0$, otherwise. In subframes configured for CSI-RS transmission, the reference signal sequence $r_l^{n_s}(m)$ is mapped to complex-valued modulation symbols $a_{kl}^p = w_\alpha r_l^{n_s}(m')$ which have been designated as reference symbols on AP_p where $m' = m + \lfloor [\max(N_{\mathrm{RB}}^{\mathrm{DL}}) - N_{\mathrm{RB}}^{\mathrm{DL}}]/2 \rfloor; m = 0, 1, \ldots, N_{\mathrm{RB}}^{\mathrm{DL}} - 1$ and $w_\alpha = 1$ for $p \in \{15, 17, 19, 21\}$ and $w_\alpha = (-1)^\alpha$ for $p \in \{16, 18, 20, 22\}$. The virtual CSI identifier in the above equation is equal to the cell identity unless configured by higher layers.

[14] In coding theory, puncturing is the process of removing some of the parity bits after encoding with an error-correction code. This has the same effect as encoding with an error-correction code with a higher rate, or less redundancy. However, with puncturing the same decoder can be used regardless of how many bits have been punctured, thus puncturing considerably increases the flexibility of the system without significantly increasing its complexity. In some cases, a predefined pattern of puncturing is used in an encoder. Then, the inverse operation, known as de-puncturing, is implemented by the decoder. Puncturing is used in UMTS and LTE during the rate-matching process. Puncturing is often used with the soft decoding and Viterbi algorithm [27,40].

Samsung Ex. 1005, Page 538 of 1147

The following parameters for CSI-RS are configured via higher layer signaling [5]:

- Number of CSI-RS ports
- CSI-RS configuration
- CSI-RS subframe configuration $I_{\text{CSI-RS}}$
- Subframe configuration period $T_{\text{CSI-RS}}$
- Subframe offset $T_{\text{Offset-CSI-RS}}$
- UE assumption on reference PDSCH transmitted power for CSI feedback P_c. The parameter P_c is the assumed ratio of PDSCH energy per resource element (EPRE)[15] to CSI-RS EPRE when UE derives CSI feedback and takes values in the range of [−8 dB to 15 dB] in 1 dB step sizes.

Multiple CSI-RS configurations can be used in a given cell, which include zero or one configuration for which the UE can assume non-zero transmission power, and zero or more configurations for which the UE can assume zero transmission power. The UE assumes that CSI-RSs are not transmitted in the special subframe(s) in the case of frame structure type 2, in subframes where transmission of a CSI-RS would collide with transmission of synchronization signals, PBCH, or system information block type 1 messages, in subframes configured for transmission of paging messages for any UE with the cell-specific paging configuration. The resource elements (k, l) that are used for transmission of CSI-RS on any of the antenna ports $AP_p = \{15\}, \{15, 16\}, \{17, 18\}, \{19, 20\}, \{21, 22\}$ are not used for transmission of PDSCH on any antenna port in the same slot, and are not used for CSI-RS on any antenna port other than those in the above set of antennas in the same slot. The mapping of CSI-RS for 1, 2, 4, and 8 antenna ports are shown in Figure 9.58 [3].

Depending on the number of CSI-RS antenna ports, there are multiple reuse patterns on different locations which share a common base pattern. In the case of 1, 2, 4, and 8 antenna ports, there are 40, 20, 10, and 5 reuse patterns, respectively, allowing different cells to utilize different reuse patterns to avoid inter-cell CSI-RS collision. In addition, the base patterns for different numbers of CSI-RS antenna ports have a nested structure, allowing for simpler CSI-RS transmitter and receiver implementation. Each antenna port's CSI-RS is spread over two adjacent resource elements across time such that two different CSI-RSs are code-

[15]Downlink power control determines the energy per resource element (EPRE). The term resource element energy denotes the energy prior to cyclic prefix insertion. The EPRE is calculated by averaging the energy over all constellation points based on the modulation scheme applied. A UE may assume downlink cell-specific reference signal EPRE is constant across the downlink system bandwidth and constant across all subframes until different cell-specific RS power information is received. The downlink cell-specific reference-signal EPRE can be derived from the downlink reference-signal transmit-power information provided by higher layers. The downlink reference-signal transmit power is defined as the linear average over the power contributions in Watts of all resource elements that carry cell-specific reference signals within the operating system bandwidth [5].

division multiplexed over two adjacent resource elements. One of the advantages of code-division-multiplexing in this case, relative to other multiplexing schemes such as FDM, is that CDM can balance transmission power across antenna ports in the frequency-domain. The CSI-RS density is related to the channel estimation accuracy. In general, larger CSI-RS density may provide better CSI measurement accuracy while reducing downlink resource utilization. Therefore, adequate CSI-RS density, where the CSI-RS occupies minimum downlink resources while providing reasonable CSI measurement accuracy, is desirable. Since CSI-RS is only used for downlink CSI measurement, its granularity is coarse due to feedback overhead; the system performance is relatively insensitive to the channel measurement accuracy as compared to that for DM-RS. Therefore, a reduced-overhead CSI-RS transmitted only once every 5, 10, 20, 40, or 80 ms is utilized in the LTE-Advanced. A typical transmission of CSI-RS, which may consist of four antenna ports with periodicity of 10 ms, would only require an overhead equivalent to 0.2% of the entire time and frequency resources. The low overhead is achieved by allocating a single resource-element-pair per CSI-RS antenna port, except when the CSI-RS has only one antenna port, where two resource elements are allocated per resource block pair (Table 9.12). The configuration period $T_{\text{CSI-RS}}$ and the subframe offset $T_{\text{Offset-CSI-RS}}$ for CSI-RSs are given in Table 9.13. The parameter $I_{\text{CSI-RS}}$ can be configured separately for CSI-RSs for which the UE assumes non-zero and zero transmission power. Those subframes containing CSI-RSs must satisfy the following equation $(10n_{\text{subframe}} + \lfloor n_s/2 \rfloor - T_{\text{Offset-CSI-RS}}) \bmod T_{\text{CSI-RS}} = 0$ [5,14].

Although different CSI-RS patterns are likely to be used in neighboring cells to minimize CSI-RS collision, neighbor cells' downlink data signals may still significantly interfere with CSI-RS received at a cell-edge UE in a heavily loaded network in interference-limited scenarios. In addition, accurate CSI measurement for neighbor cells is also required for support of coordinated multipoint transmissions. To mitigate the interference effects and to facilitate CSI measurement of neighbor cells, LTE-Advanced supports data reference element muting, whereby data reference elements colliding with CSI-RS in a neighbor cell are transmitted with zero transmission power (i.e., muted resource elements).

Table 9.12 Relative Overhead of CSI-RS

Number of Antenna Ports	1	2	4	8
Cell-specific reference signal	4.76	9.52	14.28	–
UE-specific reference signal	7.14	7.14	14.28	14.28
CSI reference signal	$1.19/T_{\text{CSI-RS}}$	$1.19/T_{\text{CSI-RS}}$	$2.38/T_{\text{CSI-RS}}$	$4.76/T_{\text{CSI-RS}}$

Note that CSI-RS overhead is a function of CSI-RS periodicity $T_{\text{CSI-RS}}$ in number of subframes.

Table 9.13 CSI Reference Signal Subframe Configuration [5]

CSI-RS Subframe Configuration Parameter $I_{CSI\text{-}RS}$	CSI-RS Periodicity $T_{CSI\text{-}RS}$ (Subframes)	CSI-RS Subframe Offset $T_{Offset\text{-}CSI\text{-}RS}$
0–4	5	$I_{CSI\text{-}RS}$
5–14	10	$I_{CSI\text{-}RS} - 5$
15–34	20	$I_{CSI\text{-}RS} - 15$
35–74	40	$I_{CSI\text{-}RS} - 35$
75–154	80	$I_{CSI\text{-}RS} - 75$

For an LTE-Advanced UE, the eNB applies code rate-matching around the muted resource elements so that the LTE-Advanced UE is able to decode the allocated data sub-carriers.

As mentioned earlier, a cell can be configured with one, two, four, or eight antenna-port CSI-RS. The exact CSI-RS structure, including the exact set of resource elements used for CSI-RS in a resource block, depends on the number of antenna ports for which the CSI-RS is configured within the cell and may vary for different cells. There are 40 possible positions within a resource-block pair that the CSI-RS can be mapped, and in a given cell, a subset of the corresponding resource elements is used for CSI-RS transmission. In the case of two antenna-port CSI-RS within a cell, the CSI-RS consists of two adjacent reference symbols per resource-block pair. The CSI-RS pair is separated by applying OCCs to the reference-symbol pair. Thus, there can be 20 different CSI-RS configurations based on the 40 possible reference-symbol positions.

In the case of four or eight antenna-port CSI-RS, the CSI-RS are pairwise frequency multiplexed. Thus, there is a possibility for ten or five different antenna ports for which the CSI-RS is configurations, each supporting four or eight CSI-RS, respectively. In the case of a single antenna-port CSI-RS, the same structure as for two antenna-port CSI-RS is used, although with only one of the OCCs. The overhead of a single antenna-port CSI-RS, in terms of occupied resource elements, is the same as two antenna-port CSI-RSs. In the time-domain, the CSI-RS can be transmitted with different periods, ranging from 5 to 80 ms. With a 5 ms period, the overhead per CSI-RS is roughly 0.12% per CSI-RS (one resource element per CSI-RS per resource-block pair, in every fifth subframe), thus longer periods result in lower overhead. The exact subframe in which CSI-RS is transmitted in a cell can also be configured, allowing for separation of CSI-RS transmissions among cells in the time-domain, in addition to using different sets of resource elements within the same subframe. In subframes in which CSI-RS is transmitted, the CSI-RS is transmitted in every resource block across the frequency. In other words, the CSI-RS transmission covers the entire cell bandwidth and by this means it is considered a wideband reference signal.

As mentioned above, when DM-RSs are transmitted within a resource block, the corresponding resource elements on which the reference signals are transmitted are explicitly avoided when mapping PDSCH symbols to the resource block. This restricted resource mapping, which is obviously not understood by legacy UEs, is possible as DM-RSs can be assumed to be transmitted only in resource blocks in which terminals supporting such reference signals are scheduled. In general, a legacy terminal is assumed to be never scheduled in a resource block where DM-RSs are transmitted. The situation is somewhat different for CSI-RS. As CSI-RS is transmitted within all resource blocks in the frequency-domain, it would not be practically possible to assume that legacy terminals are never scheduled in a resource block in which CSI-RS is transmitted. If the PDSCH mapping were modified to explicitly avoid the resource elements in which CSI-RS is transmitted, the mapping would not be recognized by legacy terminals. In order to alleviate this problem, the PDSCH is mapped consistent with LTE Rel-8, i.e., the mapping is not modified to avoid the resource elements on which CSI-RS is transmitted, and CSI-RS is transmitted on top of the corresponding PDSCH symbols. This will obviously impact the PDSCH demodulation performance, as some PDSCH symbols will be interfered by CSI-RS transmission. However, the remaining PDSCH symbols will not be impacted and the PDSCH will still be decodable with some acceptable performance degradation. On the other hand, if a Rel-10 terminal is scheduled in a resource block in which CSI-RS is transmitted, the PDSCH mapping is modified to explicitly avoid the resource elements on which the CSI-RS is transmitted. The remaining resource elements are used for PDSCH transmission within the cell. However, it is also possible to additionally configure one or several subsets of the resource elements as muted CSI-RS.

A muted CSI-RS has the same structure as an unmuted CSI-RS, except that nothing is actually transmitted on the corresponding resource elements. A muted CSI-RS can thus be seen as a normal CSI-RS with zero power. The motivation for inclusion of the muted CSI-RS is to enable creation of transmission holes corresponding to actual CSI-RS transmissions in neighboring cells in order to: (1) make it possible for a terminal to receive CSI-RS of neighboring cells, without being severely interfered by transmissions in its serving cell. The CSI estimation on neighboring cells would be necessary for multi-cell-transmission techniques and (2) to reduce the interference to CSI-RS transmissions in other cells. This is especially applicable to heterogeneous networks where overlapping layers with cells of substantially different transmit power are expected to coexist in the same frequency band. There can be one or multiple sets of muted CSI-RS in a cell. In the case of interference avoidance to neighboring cells, it is typically sufficient to use a single set of muted CSI-RS, as one should typically avoid interference to a single lower-power cell. When muted CSI-RS are used to alleviate intra-cell interference, in order to improve reception of CSI-RS from neighboring cells, multiple sets of muted CSI-RS are needed since neighboring cells may use different CSI-RS configurations.

9.10.6 Enhanced PDCCH demodulation reference signals

The LTE Rel-8 PDCCH was designed with the underlying assumption that each eNB with collocated antennas serves each cell in the network and different cells are distinguished through different physical cell identifiers. Furthermore, the UE relies on CRSs to detect and decode PDCCH content. However, the increasing mobile data traffic necessitates the use of dense and heterogeneous networks wherein multiple low-power remote radio heads (RRHs) with geographically displaced antennas share the same cell identity with the overlaying macro eNB. The cell capacity may be dramatically increased due to spatially split gain of RRHs. In such cases, if one common PDCCH region is shared among the macro eNB and all RRHs, the time-division-multiplexed PDCCH suffers from limited capacity when confined to three OFDM symbols. Moreover, since the legacy interleaved PDCCH spans across the entire system bandwidth, frequency-domain inter-cell interference coordination schemes for control channels transmitted by different RRHs cannot be utilized by PDCCH.

A Rel-11 work item initiated the design of a new downlink control channel structure referred to as ePDCCH which relies on DM-RSs for coherent detection and is frequency-division-multiplexed with data region. The use of DM-RSs would allow precoding and beamforming of the downlink physical control channels in the future. Figure 9.59 illustrates the structure of ePDCCH in a subframe and how it can coexist with the legacy PDCCH without affecting the interoperability with the legacy LTE systems [178]. As shown in the figure, the ePDCCH regions are time-division-multiplexed with the legacy PDCCH region, are frequency-division-multiplexed with PDSCH, and avoid other essential physical channels and signals. Since ePDCCH only relies on UE-specific reference signals for demodulation, the use of ePDCCH in MBSFN subframes would eliminate the reliance on CRS for control channel demodulation in the future releases of LTE.

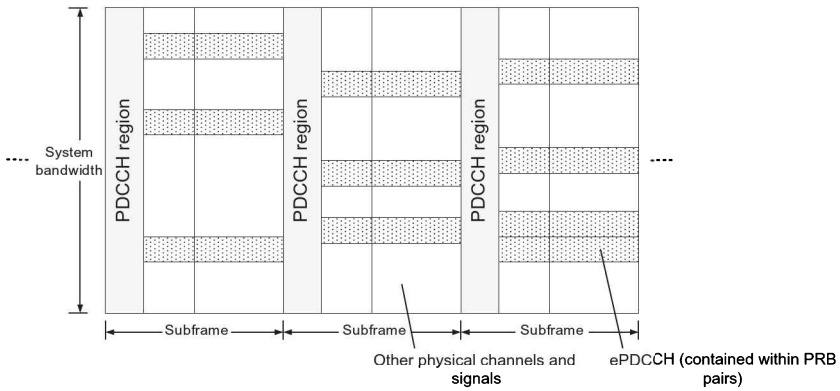


FIGURE 9.59

Enhanced PDCCH structure [149–155].

Each ePDCCH may contain multiple enhanced control channel elements (eCCEs), and multiple eCCEs can be aggregated to achieve different coding rates which depend on the channel conditions of the UE to guarantee a block error rate (BLER) of less than 1%. The existing structure of the PDCCH relies on CRS for channel estimation and coherent detection. It is shown that CRS has more overhead and is less effective when using closed-loop precoding techniques, beamforming, and MU-MIMO. The beamforming of the control channels can significantly improve the performance of the cell-edge users. The use of DM-RS or UE-specific reference signals instead of CRS would facilitate the use of beamforming for the control channels. Figure 9.60 illustrates a hypothetical beam-formed ePDCCH and depicts the localized and distributed ePDCCH mapping to the resource elements and antenna ports [149–155].

The ePDCCH DM-RSs are transmitted on antenna ports $p \in \{107, 108, 109, 110\}$ and are used as a reference for ePDCCH coherent demodulation. An ePDCCH DM-RS is not transmitted over resource element (k, l), where k denotes the frequency index and l denotes the OFDM symbol index within a slot, when one of the physical channels or physical signals are transmitted using the same resource element regardless of their antenna port. For the antenna port $p \in \{107, 108, 109, 110\}$, the reference-signal sequence $r(m)$ is defined by the equation $r(m) = (1/\sqrt{2})(1 - 2c(2m)) + j(1/\sqrt{2})(1 - 2c(2m + 1))$, where $m = 0, 1, \ldots,$

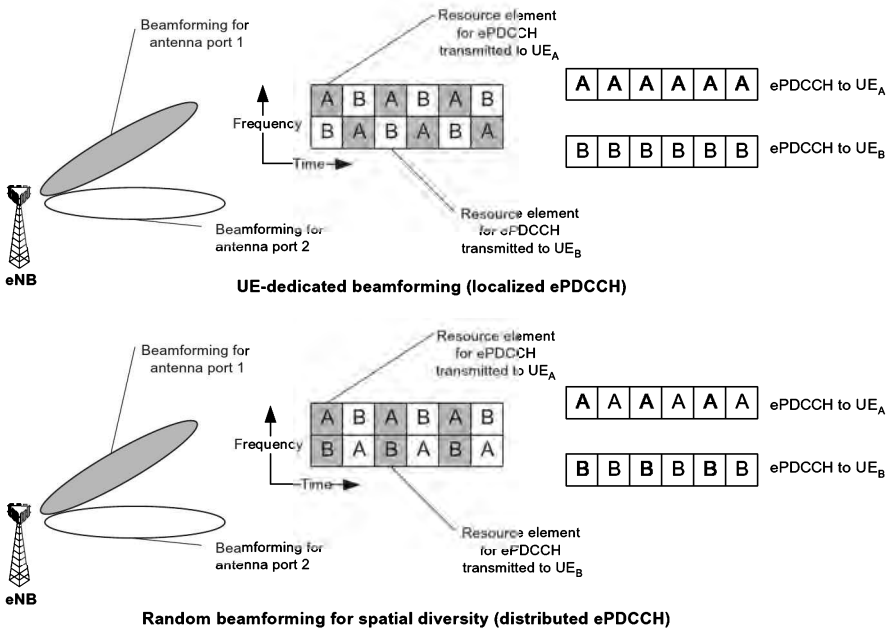


FIGURE 9.60

Possible use of beamforming for transmission of ePDCCH [149–155].

$12\max(N_{RB}^{DL}) - 1$ for a normal cyclic prefix, and $m = 0, 1, \ldots, 16\max(N_{RB}^{DL}) - 1$ for an extended cyclic prefix. The pseudo-random sequence $c(i)$ was defined in the previous sections and is initialized with the initial value $c_{init} = \left(\lfloor n_s/2 \rfloor + 1\right)\left(2N_{ID}^{ePDCCH(i)} + 1\right)2^{16} + N_{SCID}^{ePDCCH}$ at the start of each subframe where $N_{SCID}^{ePDCCH} = 2$ and $N_{ID}^{ePDCCH(i)}|_{i=0,1}$ is configured by higher layers, and index i denotes the set of ePDCCH associated with the DM-RS. For the antenna port $p \in \{107, 108, 109, 110\}$ in a PRB n_{PRB} assigned to the associated ePDCCH, the reference signal sequence $r(m)$ is mapped to complex-valued modulation symbols a_{kl}^{p} in a subframe according to the following mapping rule $a_{kl}^{p} = w_p(l')r(3l'\max(N_{RB}^{DL}) + 3n_{PRB} + m')$ for a normal cyclic prefix, where $w_p(i) = \overline{w}_p(i)$, if $(m' + n_{PRB}) \bmod 2 = 0$, and $w_p(i) = \overline{w}_p(3 - i)$ if $(m' + n_{PRB}) \bmod 2 = 1$ and the frequency index is given by $k = 5m' + 12n_{PRB} + k'$. The length-4 orthogonal cover sequence $\overline{w}_p(i)$ is defined in Table 9.14.

For an extended cyclic prefix, ePDCCH DM-RSs are not supported on antenna ports $p = 109$ and $p = 110$. The resource elements (k, l) used for transmission of ePDCCH DM-RSs to a single UE on any antenna ports in the set $S = \{107, 108\}$ or $S = \{109, 110\}$ is not used for transmission of ePDCCH on any other antenna port in the same slot, and further it is not used for DM-RSs to the same UE on any antenna port other than those in set S in the same slot. The mapping of the ePDCCH DM-RSs to resource elements within a UE-allocated resource block is identical to that corresponding to UE-specific reference signals on antenna port numbers 7–10 when replacing the antenna port numbers with 107–110 in Figure 9.57 for a normal cyclic prefix [3].

9.10.7 Positioning reference signals

Positioning is defined as the process of determining the location and/or velocity of a device using radio signals. The positioning reference signals (PRS) were introduced in LTE Rel-9 on antenna port 6, after the cell-specific reference signals were not found suitable for positioning due to the fact that the required probability of detection could not be guaranteed. A neighbor cell with its synchronization signals (primary/secondary synchronization signals) and reference

Table 9.14 The Sequence $\overline{w}_p(i)$ for Normal Cyclic Prefix [3]

Antenna Port p	$[\overline{w}_p(0)$	$\overline{w}_p(1)$	$\overline{w}_p(2)$	$\overline{w}_p(3)]$
107	[+1	+1	+1	+1]
108	[+1	−1	+1	−1]
109	[+1	+1	+1	+1]
110	[+1	−1	+1	−1]

signals is seen as detectable, when the SINR is at least −6 dB [121,179]. Simulations during development of Rel-9 have shown that this can be only guaranteed for 70% of all cases for the third best-detected cell, which means the second best neighboring cell by assuming an interference-free environment, which cannot be ensured in a realistic scenario. The PRS design has some similarities with cell-specific reference signals as defined in LTE Rel-8. The PRS is a pseudo-random QPSK sequence that is mapped to resource elements in diagonal patterns with shifts in frequency and time to avoid collision with CRSs, and to avoid overlapping with the control channels.

The PRS is defined by its bandwidth N_{RB}^{PRS} in terms of number of resource blocks used for PRS, time offset ($T_{\text{offset-PRS}}$), duration $N_{\text{subframe}}^{PRS}$ in terms of number of consecutive subframes, and its periodicity T_{PRS}. It must be noted that the PRS bandwidth is always smaller than the actual system bandwidth ($N_{RB}^{PRS} < \max(N_{RB}^{DL})$). The PRS is always mapped around the carrier center frequency and the unused DC sub-carrier in the downlink. In a subframe where PRS is configured, typically no PDSCH is transmitted. The positioning reference symbols of a certain cell can be configured to correspond to empty resource elements of the neighboring cells, thus enabling detection of the PRSs with satisfactory SIR values when receiving neighbor-cell PRSs [121].

Positioning reference signals are only transmitted in resource blocks over downlink subframes that are configured for PRS transmission. If both normal and MBSFN subframes are configured as positioning subframes within a cell, the OFDM symbols in the MBSFN subframe use the same cyclic prefix type as the first subframe in the frame. However, if MBSFN subframes are only configured to contain PRSs in a cell, the OFDM symbols configured for PRSs in the MBSFN region of these subframes use an extended cyclic prefix type. In a subframe configured for PRS transmission, the starting positions of the OFDM symbols configured for PRS transmission is identical to those in a subframe in which all OFDM symbols have the same cyclic prefix length as the OFDM symbols configured for PRS transmission. Positioning reference signals are transmitted on antenna port 6, and are not mapped to resource elements (k, l) allocated to BCH and primary and secondary synchronization channels regardless of their associated antenna ports. PRSs are only defined for sub-carrier spacing of 15 kHz.

The reference-signal sequence $r_l^{n_s}(m)$ is defined by $r_l^{n_s}(m) = (1/\sqrt{2})[1 - 2c(2m)] + j(1/\sqrt{2})[1 - 2c(2m+1)]$, $\forall m = 0, 1, \ldots, 2\max(N_{RB}^{DL}) - 1$, where $n_s = 0, 1, \ldots, 19$ is the slot number within a radio frame, $l = 0, 1, \ldots, 6$ is the OFDM symbol number within the slot. The pseudo-random sequence $c(i)$ is defined similar to the previous sections. The pseudo-random sequence generator in this case is initialized by $c_{\text{init}} = 2^{10}[7(n_s + 1) + l + 1](2N_{ID}^{\text{cell}} + 1) + 2N_{ID}^{\text{cell}} + N_{CP}$ at the start of each OFDM symbol where $N_{CP} = 1$ for a normal cyclic prefix and $N_{CP} = 0$ for an extended cyclic prefix. The reference signal sequence $r_l^{n_s}(m)$ is mapped to complex-valued modulation symbols a_{kl}^p designated to the PRS on antenna port $p = 6$ such that $a_{kl}^p = r_l^{n_s}(m')$ where in the case of a normal cyclic prefix

Table 9.15 Positioning Reference Signal Subframe Configuration Parameters [3]

PRS Configuration Index I_{PRS}	PRS Periodicity T_{PRS} (Subframes)	PRS Subframe Offset $T_{offset-PRS}$ (Subframes)
0–159	160	I_{PRS}
160–479	320	$I_{PRS} - 160$
480–1119	640	$I_{PRS} - 480$
1120–2399	1280	$I_{PRS} - 1120$
2400–4095		Reserved

$m' = m + \max(N_{RB}^{DL}) - N_{RB}^{PRS}$. The bandwidth for PRSs and N_{RB}^{PRS} is configured by higher layers and the cell-specific frequency shift is given by $\nu_{shift} = N_{ID}^{cell} \bmod 6$.

The cell-specific subframe configuration period T_{PRS} and the cell-specific subframe offset $T_{offset-PRS}$ for transmission of PRSs are given in Table 9.15 (see Figure 9.61). The PRS configuration index I_{PRS} is configured by higher layers. Positioning reference signals are transmitted only in configured downlink subframes and, in the case of TDD duplex scheme, are not transmitted in special subframes. The PRS is transmitted in N_{PRS} consecutive subframes that are configured by higher layers. The PRS location in time is given by $(10n_{subframe} + \lfloor n_s/2 \rfloor - T_{offset-PRS}) \bmod T_{PRS} = 0$. A UE may assume that downlink PRS EPRE is constant across the PRS bandwidth and across all OFDM symbols that contain positioning reference signals in a given PRS occurrence. The structure of the positioning reference signals in the time- and frequency-domain is illustrated in Figure 9.61.

9.10.8 MBSFN reference signals

Support of MBMS in LTE has certain impacts on the physical layer design. It starts with the use of the cyclic prefix, since it is essential for the OFDM system that multipath versions of the signal arrive at the receiver within the cyclic prefix to avoid ISI. As we have multiple, but synchronized transmission from different broadcast/multicast sources, the expected delay spread is much larger than that for unicast operation. Therefore, in order to support multicast and broadcast services, MBSFN uses an extended cyclic prefix while the majority of today's LTE networks are based on a normal cyclic prefix. In the case of the mixed (or time-division-multiplexed) MBSFN and unicast operation, the subframes that are supposed to carry MBMS content are divided into a non-MBSFN and an MBSFN region. The non-MBSFN region can occupy one or two OFDM symbols at the beginning of the subframe. In this region, PCFICH, PDCCH, and PHICH will be mapped to ensure interoperability with user terminals that do not support MBSFN. For proper operation of the network, it is required to schedule terminals to receive or transmit data, page them, or provide feedback on their recent uplink transmission. In this region, CRSs and a normal cyclic prefix will be used. The MBSFN-region will carry PMCH. The extended cyclic prefix and sub-carrier spacing of

Samsung Ex. 1005, Page 547 of 1147

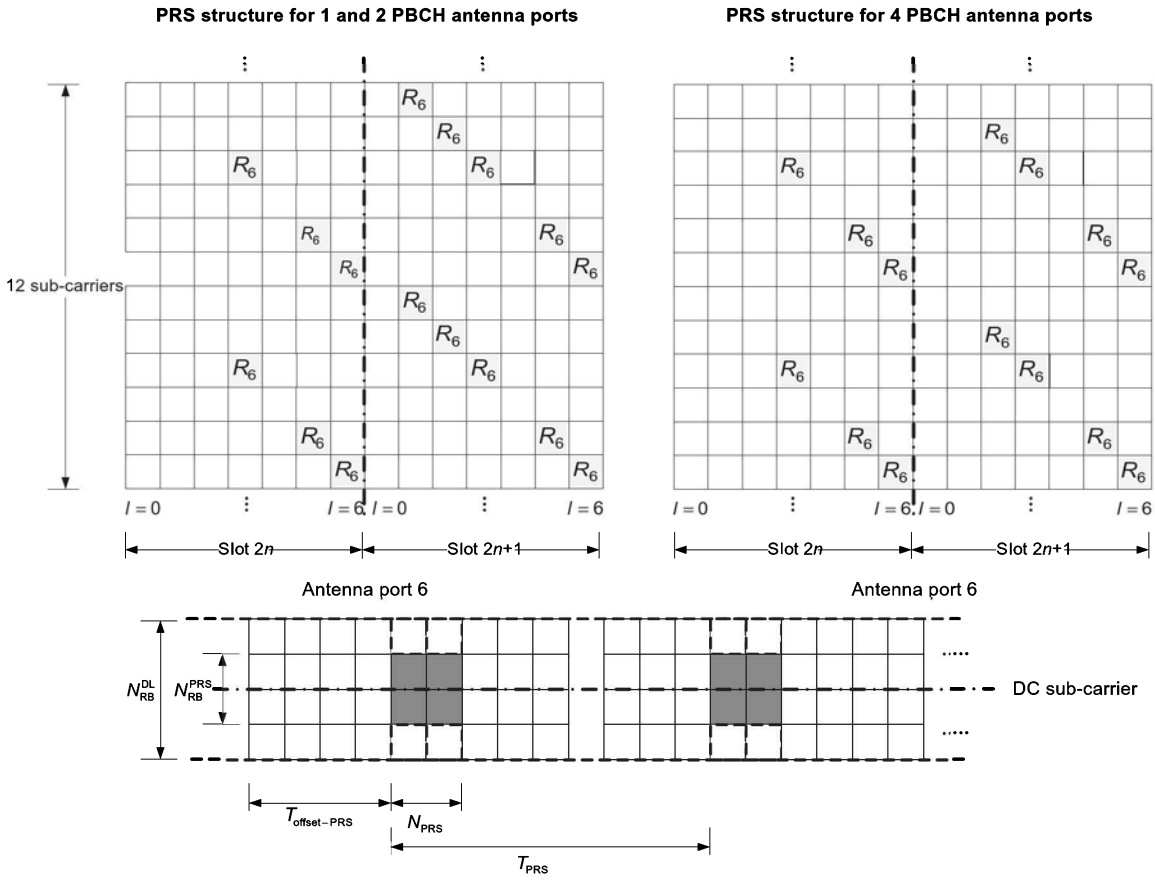


FIGURE 9.61

Structure of PRS in time- and frequency-domain [3,179].

15 kHz (or 7.5 kHz) are used in this region in order to contain the time difference of the cells that belong to an MBSFN area, while transmitting the same content to the terminals. In order to enable coherent demodulation at the terminal, as well as proper channel estimation, the use of CRSs is not sufficient. Thus new reference signals for MBMS transmission have been designed as part of Rel-9. Each cell belonging to the MBSFN area will transmit the same MBSFN reference signal pattern at the exact same time–frequency position. The MBSFN reference signals have a tighter spacing in the frequency-domain due to the frequency-selective nature of the radio channel. In contrast to the CRSs, the initialization sequence for the MBSFN reference signal does not depend on the physical cell identity, rather on the MBSFN identity provided by system information block type 13 [179].

The MBSFN reference signals are transmitted on antenna port 4 in the MBSFN region of MBSFN subframes, only if PMCH is transmitted. These reference signals are only defined for an extended cyclic prefix. The MBSFN reference-signal sequence $r_l^{n_s}(m)$ is obtained as $r_l^{n_s}(m) = (1/\sqrt{2})[1 - 2c(2m)] + j(1/\sqrt{2})\ [1 - 2c(2m+1)]$, $\forall m = 0, 1, \ldots, 6\max(N_{RB}^{DL}) - 1$, where $n_s = 0, 1, \ldots, 19$ is the slot number within a radio frame, $l = 0, 1, \ldots, 5$ is the OFDM symbol number within the slot, and $c(i)$ is the pseudo-random sequence defined similar to the previous sections, which is initialized by $c_{init} = 2^9[7(n_s + 1) + l + 1](2N_{ID}^{MBSFN} + 1) + N_{ID}^{MBSFN}$ at the start of each OFDM symbol. The reference-signal sequence $r_l^{n_s}(m')$ in OFDM symbol l is mapped to complex-valued modulation symbols a_{kl}^p on antenna port $p = 4$ according to $a_{kl}^p = r_l^{n_s}(m')$, where $m' = m + 3(\max(N_{RB}^{DL}) - N_{RB}^{DL})$. Figure 9.62 illustrates the resource elements used for MBSFN reference signal transmission.

9.10.9 Reference signal overhead calculation

The reference signals regardless of their types add to the overhead and reduce the number of available resource elements in a subframe for user data allocation. The amount of overhead typically increases when multiple transmit antennas and/or an extended cyclic prefix is used. Among various types of reference signals used in LTE/LTE-Advanced, the cell-specific reference signals correspond to the highest amount of overhead, as they are always on and span the entire system bandwidth. The overhead concerning other types of reference signals varies depending on the number of antenna ports, periodicity of the configured reference signals, number of UE-allocated resource blocks, etc. The overhead is calculated as the number of resource elements in each resource block pair that is designated to the reference signals relative to the total number of resource elements in each resource block pair. The total number of resource elements is also a function of the cyclic prefix type. Table 9.16 provides the overhead figures associated with each reference signal type for FDD systems. The overhead of the reference signals in TDD systems is also a function of the downlink/uplink subframe configuration and the number of OFDM symbols in the DwPTS portion of the special subframes [13].

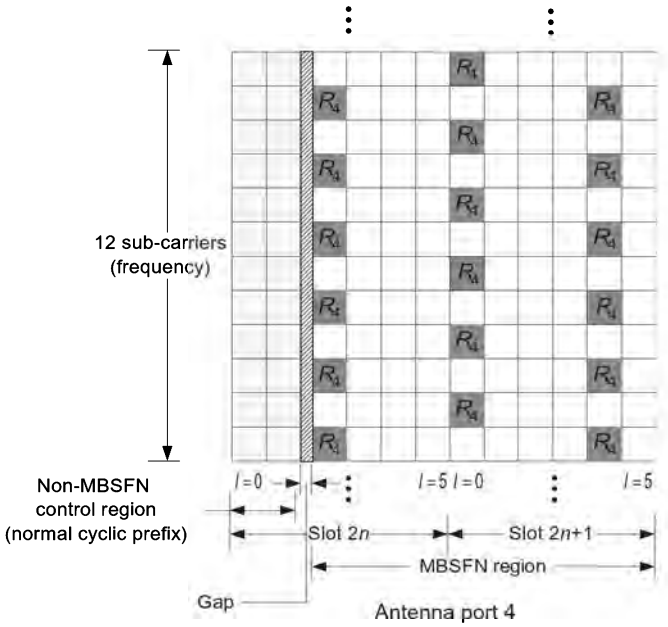


FIGURE 9.62

MBSFN reference signal structure [3,13].

9.11 Channel coding and modulation

Forward error correction (FEC) channel coding schemes are commonly used to improve the efficiency and robustness of wireless digital communication systems. On the transmitter side, an FEC encoder adds redundancy to the input data in the form of parity bits, while at the receiver, the redundancy is utilized by the FEC decoder to correct a number of channel errors. The use of FEC schemes would allow tolerance of more channel errors than otherwise can be tolerated without FEC. The use of FEC schemes enables the wireless communication systems to operate with a lower transmit power, to transmit over longer distances, to tolerate more interference, and to transmit at a higher data rate. A binary FEC encoder receives k bits at a time and produces a codeword of n bits, where $n > k$. While there are 2^n possible sequences of n bits, only 2^k codewords are used. The ratio k/n is the code rate and is denoted by R. Lower code rates corresponding to small values of R can generally correct more channel errors than higher code rates, and therefore are more robust. However, higher code rates are more bandwidth efficient because of lower overhead due to parity or redundancy bits. Thus, the selection of the code rate is a trade-off between energy efficiency and bandwidth efficiency. For every combination of code rate R, code word length n, baseband modulation scheme, channel type, and receiver noise power, there is a theoretical

Table 9.16 Reference Signal Overhead in FDD Mode [13]

Release	Type of Reference Signal	Number of Antenna Ports	Antenna Ports	Number of Reference Signals per Resource Block Pair	Overhead (%) Normal Cyclic Prefix	Overhead (%) Extended Cyclic Prefix
8	Cell-specific reference signals	1	0	8	4.8	5.6
		2	0,1	16	9.5	11.1
		4	0,1,2,3	24	14.3	16.7
	UE-specific (demodulation) reference signals	1	5	12	7.1	8.3
	MBSFN reference signals	1	4	18	–	12.5
9	Positioning reference signals	1	6	16 (14)	9.5	9.7
	UE-specific (demodulation) Reference Signals	2	7,8	12	7.1	8.3
10	UE-specific (demodulation) reference signals	1	7	12	7.1	8.3
		2	7,8	12	7.1	8.3
		4	7–10	24	14.3	16.6
		8	7–14	24	14.3	16.6
	CSI reference signals	1	15	2	$1.19/T_{\text{CSI-RS}}$	$1.39/T_{\text{CSI-RS}}$
		2	15,16	2	$1.19/T_{\text{CSI-RS}}$	$1.39/T_{\text{CSI-RS}}$
		4	15–18	4	$2.38/T_{\text{CSI-RS}}$	$2.78/T_{\text{CSI-RS}}$
		8	15–22	4	$2.38/T_{\text{CSI-RS}}$	$2.78/T_{\text{CSI-RS}}$
11	ePDCCH demodulation reference signals	4	107–110	12(24)	7.1–14.3	8.3–16.6

lower limit on the amount of energy that must be used to convey one bit of information. This limit is called the channel capacity of the communication channel [33]. Since the dawn of information theory, engineers and mathematicians have tried to construct codes that achieve performance close to Shannon capacity [27,33,39]. The introduction of convolutional turbo codes (CTCs) in 1993 followed by the invention of block turbo codes (BTCs) in 1994 closed much of the performance gap with the maximum channel capacity. Turbo codes are essentially parallel concatenated convolutional codes (PCCCs) with an internal interleaving mechanism combined with an iterative soft-decision decoding algorithm. In this section, we first review the principles of convolutional and turbo coding, and then we describe the channel coding and modulation schemes that are used in LTE/LTE-Advanced systems.

Before we proceed to the next sections, we need to clarify that detection, decision-making, hypothesis testing, and decoding are synonymous. The word detection refers to the effort to decide whether some phenomenon is present or not on the basis of some observations. For example, a radar system uses the data to detect whether a target is present. The meaning has been extended in the communication field to detect which one, among a set of mutually exclusive alternatives, is correct. Decision-making is the process of deciding between a number of mutually exclusive alternatives. Hypothesis testing is the same, and here the mutually exclusive alternatives are called hypotheses. Decoding is the process of mapping the received signal into one of the possible sets of codewords or transmitted symbols. We use the word hypotheses for the possible choices in what follows, since the word conjures up the appropriate intuitive image [100–108].

9.11.1 Principles of convolutional coding

Convolutional codes have been extensively used in telecommunication systems such as digital video, radio, mobile, and satellite communication in order to achieve reliable data communication [27]. Prior to the invention of the turbo codes, convolutional code was the most efficient coding scheme that provided the closest performance to the Shannon limit. A generic convolutional encoder (k, n, K) is characterized by a shift register with K stages, k input bits, and n output bits. The rate of the convolutional code is defined as $R = k/n$. A convolutional encoder is a discrete linear time-invariant system and can be viewed as a finite-state machine with 2^{K-1} states. Therefore, convolutional codes are linear codes which do not divide the source stream into blocks, but instead read and transmit bits continuously. The transmitted bits are a linear function of the past source bits. Usually the rule for generating the transmitted bits involves feeding the present source bit into a linear-feedback shift-register of length K, and transmitting one or more linear functions of the state of the shift register in each iteration. The resulting transmitted bit stream is the convolution of the source stream with a linear filter (Figure 9.63). The impulse–response function of this filter may have finite or infinite duration depending on the choice of feedback shift-register,

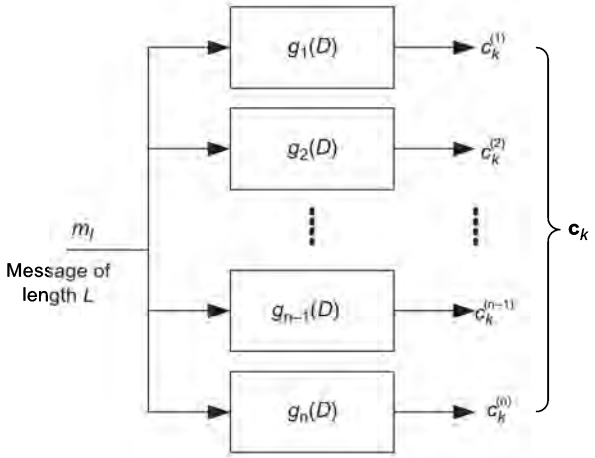


FIGURE 9.63

General form of (1/*n*) convolutional coder [27,40].

i.e., recursive or non-recursive convolutional codes. As an example, consider the LTE convolutional coder with $k = 1, n = 3, K = 6$, and $R = 1/3$. The linear combinations are defined via $n = 3$ generator sequences where $\mathbf{G} = \{\mathbf{g}_0, \mathbf{g}_1, \ldots, \mathbf{g}_{n-1}\}$ and $\mathbf{g}_i = \{g_{i0}, g_{i1}, \ldots, g_{i(K-1)}\}$. The generator sequences used in Rel-8 of LTE are given by $\mathbf{g}_0 = \{1\ 0\ 1\ 1\ 0\ 1\ 1\}$, $\mathbf{g}_1 = \{1\ 1\ 1\ 1\ 0\ 0\ 1\}$, and $\mathbf{g}_2 = \{1\ 1\ 1\ 0\ 1\ 0\ 1\}$ as shown in Figure 9.63. Puncturing is used to make an arbitrary k/n rate code from a basic rate 1/2 or 1/3 code by deleting some bits in the encoder output. For example, if we want to make a code with rate 2/3, we should take the basic 1/2-encoder output and transmit every second bit from the first branch and every bit from the second branch.

Traditional channel coding techniques achieve coding gain at the expense of bandwidth expansion, where coding gain is defined as the reduction in required E_b/N_0 for a given error probability. In the case of band-limited channels, bandwidth expansion is usually not possible. Trellis-coded modulation schemes use redundant non-binary modulation in conjunction with a finite state encoder. For each symbol interval, the encoder selects one of a set of *M*-ary waveforms, thereby generating a sequence of coded waveforms. The noisy signals at the receiver are detected and decoded by a soft-decision maximum-likelihood (ML) detector/decoder. Coding gain can be achieved without sacrificing data rate or expanding bandwidth, but at the expense of increasing decoder complexity.

A convolutional encoder, which is an example of trellis coding, is a finite-state machine, i.e., the output is a function of both the current input and the state of the encoder. A convenient way of characterizing such a state-oriented process is through a trellis diagram. Figure 9.64 illustrates a trellis diagram for a

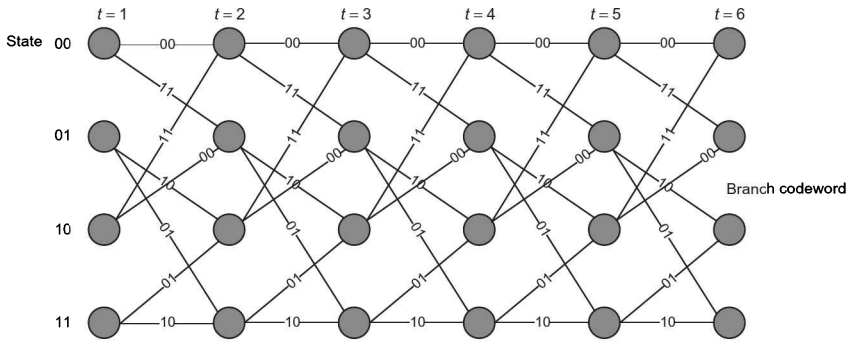


FIGURE 9.64

Four-state Trellis diagram [27,40].

hypothetical four-state convolutional encoder. The number of stages in the encoder shift register is referred to as its constraint length K, and the contents of the rightmost $K-1$ stages define the state of the encoder. Each node of the trellis represents a distinct state at a given time. A branch between any two states represents the transition between those states, and is labeled with the output code sequence associated with that transition. The most important property of the trellis diagram is that every possible sequence of message bits into the encoder corresponds to a unique path through the trellis and a unique sequence of coded bits, and vice versa. The task of the decoder is to estimate the path that the encoded message traverses through the trellis. If all input message sequences are equally likely, a decoder that will achieve the minimum probability of error is one that compares the conditional probabilities $p(\mathbf{z}_j|\mathbf{u}_m)$, where $\mathbf{z}_j$ is the received sequence and $\mathbf{u}_m$ is one of the possible transmitted sequences, and chooses the sequence corresponding to the ML. The decoder chooses $\mathbf{u}_{m'}$ if $p(\mathbf{z}_j|\mathbf{u}_{m'}) = \max_{\forall \mathbf{u}_m} p(\mathbf{z}_j|\mathbf{u}_m)$. Finding the sequence which maximizes $p(\mathbf{z}_j|\mathbf{u}_m)$ is equivalent to finding the sequence $\mathbf{u}_{m'}$ that has the greatest similarity to $\mathbf{z}_j$. Over a hard-decision channel, the similarity is often estimated by the Hamming distance between received and candidate sequences, and over a soft-decision channel it is estimated by the Euclidean distance. Since an ML decoder will choose the trellis path whose corresponding sequence $\mathbf{u}_{m'}$ is at the minimum distance to the received sequence $\mathbf{z}_j$, the ML problem is identical to the problem of finding the shortest distance through the trellis diagram [19]. A comparison of the performance of a rate 1/2 convolutional code with hard- and soft-decision decoding ($K = 7$ and BPSK modulation) over an AWGN channel is shown in Figure 9.66. The advantage of a soft-decision decoding over a hard-decision decoding convolutional coder (approximately $+2$ dB coding gain) can be seen in this example. The performance of the same encoder with hard-decision decoding under Rayleigh multipath fading and AWGN channels for different number paths has been calculated and compared in Figure 9.66.

Samsung Ex. 1005, Page 554 of 1147

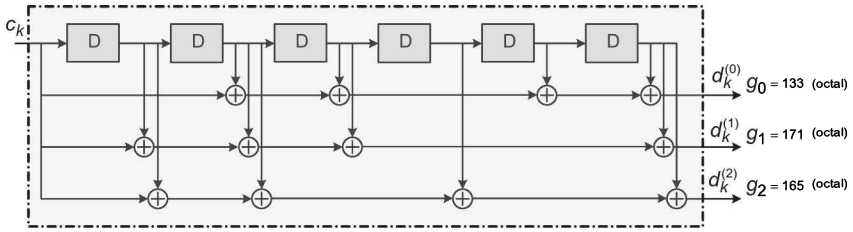


FIGURE 9.65

LTE rate 1/3 tail-biting convolutional encoder [4].

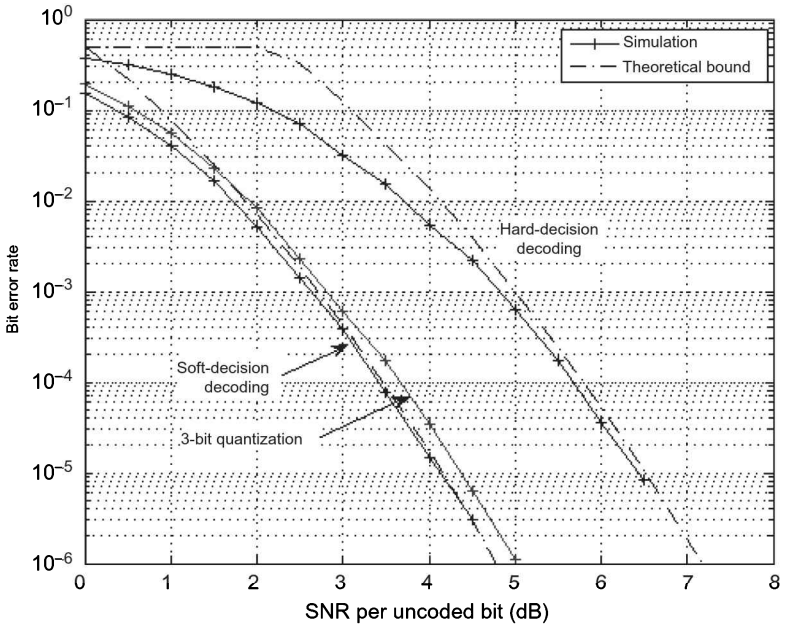


FIGURE 9.66

Performance of rate 1/2 convolutional code ($K = 7$ and BPSK modulation) over AWGN channel [141].

In order to terminate a convolutional code, some bits are appended to the information sequence such that the shift-register's content returns to the initial state. Each of the k input sequences of length L bits is padded with K zeros, and these kL input sequences jointly generate $n(K + L)$ output bits. The effective rate of the terminated convolutional code is now $R_{\text{effective}} = kL/n(K + L) < k/n$, which is less than the nominal rate due to the addition of redundancy. Thus, terminating the trellis of a convolutional code is essential in the code's performance for packet-based communications.

Samsung Ex. 1005, Page 555 of 1147

Tail-biting convolutional coding is a technique of trellis termination which avoids the rate loss incurred by zero-tail termination at the expense of a more complex decoder. Tail-biting encoding ensures that the starting state of the encoder is the same as its ending state, and that this state value does not necessarily have to be the all-zero state. For a rate $1/n$ feed-forward encoder, this is achieved by initializing the K memory elements of the encoder with the last K information bits of a block of data of length L and ignoring the output. All of the L bits are then input to the encoder and the resulting nL output bits are used as the codeword. In comparison, the zero-tail termination method appends K zeros to a block of data to ensure the feed-forward encoder starts from and ends in the all-zero state for each block. This incurs a rate loss due to the extra tail bits (i.e., non-informational bits) that are transmitted.

The ML tail-biting decoding involves determining the best path in the trellis under the constraint that it starts and ends in the same state. One way to implement this method is to run M parallel Viterbi algorithms, where M is the number of states in the trellis, and select the decoded bits based on the Viterbi algorithm that gives the best metric. However, this makes the decoding M times more complex relative to the zero-tailed encoding.

The structure of the LTE rate 1/3 tail-biting convolutional encoder with constraint length 7 shown in Figure 9.67. The bit sequence input for a given code block to channel coding is denoted by $(c_0, c_1, \ldots, c_{K-1})$, where K is the number of bits to be encoded. The initial value of the shift register (the initial state) of the encoder is set to the values corresponding to the last 6 information bits in the input stream, so that the initial and final states of the shift register are the same. Therefore, if the content of the shift register is denoted by $(s_0, s_1, s_2, s_3, s_4, s_5)$, then the initial value of the shift register is set to $s_i = c_{(K-1-i)}$. The encoder output streams $d_k^{(0)}$, $d_k^{(1)}$, and $d_k^{(2)}$ correspond to the first, second, and third parity streams [4].

The performance of an example rate 1/2 tail-biting convolution coder in an AWGN channel with different modulation orders, coding rates, and data frame sizes using a Viterbi decoder has been evaluated and shown in Figure 9.68 where the wrap depth is a predetermined number of trellis stages used as a parameter in the figure. One way to estimate the required wrap depth for MAP decoding[16] of a

[16]There are a number of possible criteria to use in making decisions, and initially, we assume that the criterion is to maximize the probability of choosing correctly. That is, when the experiment is performed, the resulting sample point maps into a sample value h for H and a sample value y for Y. The decision-maker observes y (but not h) and maps y into a decision $\hat{H}(y)$. The decision is correct if $\hat{H}(y) = h$. In principal, maximizing the probability of choosing correctly is almost trivially simple. Given y, we calculate $p_{H|Y}(h|y)$ for each $h, 0 \le h \le M-1$. This is the probability that h is the correct hypothesis conditional on y. Thus the rule for maximizing the probability of being correct is to choose $\hat{H}(y)$ to be that h for which $p_{H|Y}(h|y)$ is maximized. This is denoted as $\hat{H}(y) = \arg\max_h[p_{H|Y}(h|y)]$. If the maximum is not unique, it makes no difference to the probability of being correct which maximizing h is chosen. To be explicit, we will choose the smallest maximizing h. The conditional probability $p_{H|Y}(h|y)$ is called an *a posteriori* probability, and thus the above decision rule is called the maximum a posteriori probability (MAP) rule [45].

Samsung Ex. 1005, Page 556 of 1147

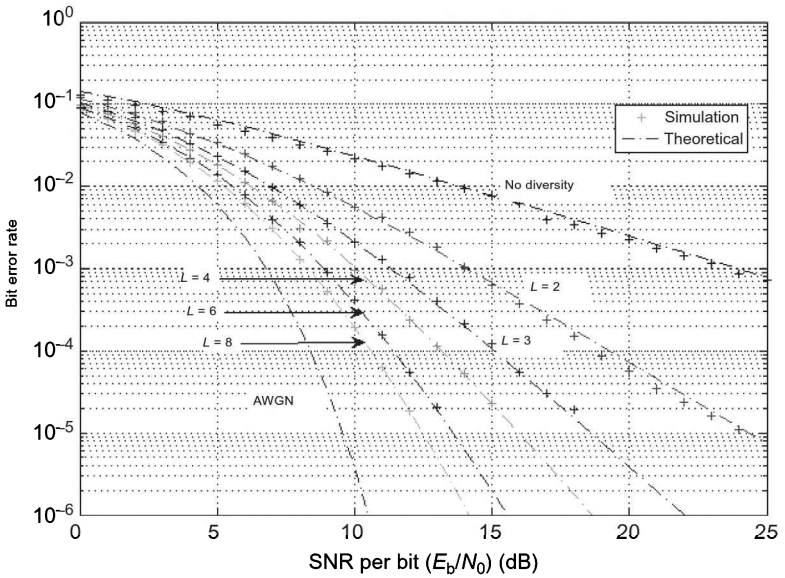


FIGURE 9.67

Performance of rate 1/2 convolutional code (K = 7 and BPSK modulation) in multi-path Rayleigh fading channel [141].

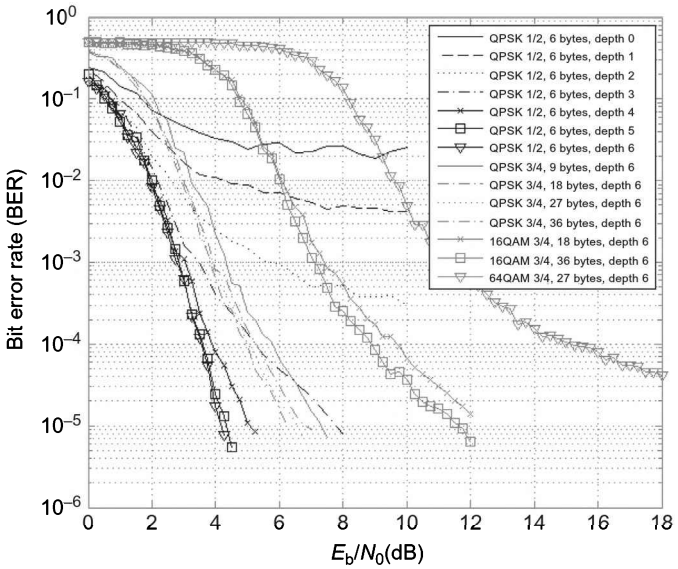


FIGURE 9.68

Performance of an example rate 1/2 tail-biting convolutional coder in an AWGN channel [144].

tail-biting convolutional code is to conduct hardware or software experiments, requiring a circular MAP decoder with a variable wrap depth to be implemented in order to measure the decoded bit error rate versus E_b/N_0 for successively increasing wrap depths. The minimum decoder wrap depth that provides the minimum probability of decoded bit error for a specified E_b/N_0 is found when further increases in wrap depth do not decrease the error probability, as shown in Figure 9.68. It can be seen that the link-level BER performance improves when increasing the depth and the performance does not significantly change after a certain limit. The effect of different data frame sizes (6, 9, 18, 27, and 36 bytes) on the BER performance has been investigated and is shown in the figure.

9.11.2 Principles of turbo coding

In information theory, one can ideally approach the Shannon limit as closely as desired using soft-decision decoding of a long block code, or a convolutional code with a large constraint length. In practice, however, decoding of these codes becomes very computationally intensive. Similar to any conventional error correcting codes, the turbo codes work by imposing a structure on the transmitted bit sequence. If the received bit sequence does not match this known structure, the receiver declares an error has occurred. If the number of errors is sufficiently small and the structure is robust, the receiver can detect the erroneous bits and reconstruct the correct bit sequence. A strong error correcting code has two key characteristics: (1) the encoder imposes a structure that maximizes the difference (or the distance) between any two valid bit sequences; and (2) the decoder utilizes all the information available at the receiver's end including the redundancy and previously unsuccessful transmissions. Thus, the decoder can determine which valid bit sequence is most likely the one that has been transmitted. The first significant difference between a turbo code and conventional code is the use of a recursive systematic encoder. A typical convolutional coder uses a non-recursive structure. By feeding one of the outputs back to the input, a recursive encoding structure is obtained. The recursive structure is systematic, i.e., the input bits appear directly as part of the encoded bit stream. Therefore, a systematic structure enables the combination of two codes in order to construct a stronger composite code. The input bit stream only needs to be transmitted once. The computed parity bits from each of the two constituent encoders are transmitted separately. The recursive structure interacts with the interleaver to give the composite code some unique performance characteristics. The optimum decoder for any code is an ML decoder. In the ML decoding, the receiver produces a probability that each received bit is either 0 or 1. The ML decoder then uses these received bit probabilities, along with its knowledge of the structure imposed by the encoder, to compute the probability of every possible transmitted bit sequence. The most probable transmitted bit sequence is then chosen as the decoded bit sequence. The ML decoding is theoretically optimal

and can achieve the lowest probability of decoder failure or probability of error. However, there is no efficient way to practically compute these probabilities when the block size is large. This leads to the use of suboptimal, but realizable, decoding methods.

The second unique feature of turbo codes is their decoder, which is based on iterative application of ML decoding. The ML decoding of a convolutional code would be easy, if the constraint length (register size) is short; however, short codes are relatively weak. Combining two such codes produces a much stronger code, but now ML decoding becomes intractable. Each shorter code can be ML decoded if we assume the other code has already been decoded. This leads to an iterative decoding strategy, which is to decode the first code and then use the resulting updated probabilities to decode the second code. Once this is done, the updated results can more accurately decode the first code. This decoding strategy shares the limitations of all iterative solution methods, including the possibility of entrapment in a local minimum. The decoder output is not necessarily the true global ML solution. But the study of the corresponding link-level BER curves suggests that it tends to be a very good solution.

In a convolutional code, the input bit sequence usually has tail bits added. This initializes the encoder state to all zeros at the end of the bit sequence. In a turbo coder, this implies that a valid bit sequence cannot contain a single one. It must either be all zeros, or it must contain two or more ones. Because encoding is a linear process, this implies that any valid input sequence must differ from any other valid input sequence in at least two bits. In this case, if we consider only the first decoder's output, it differs from the correct output in two positions. However, the turbo encoder contains an interleaver. It must be noted that the two constituent encoders are operating on the same set of bits, but in a different order. Therefore, although the error bits are close together for the first decoder, they are widely separated in the second decoder's input. As a result, the second decoder's output differs from the correct output in many bit positions. It will see the incorrect bit sequence as highly implausible. This is the basic principle of operation of the turbo codes, and is the key to the high performance of turbo codes at low SNRs. Erroneous information sequences that look reasonable to one decoder are likely to be rejected by the other decoder. This is also the rationale for the error floor at higher SNRs. Since the interleaver is random, there are a few error patterns that will look plausible to both decoders. As the SNR increases, these weak patterns come to dominate the BER curve.

One of the most interesting characteristics of a turbo code is that it is not just a single code. It is, in fact, a combination of two codes that work together to achieve a synergy that would not be possible by merely using one code by itself. As shown in Figure 9.69, a turbo code is formed from the parallel concatenation of two constituent encoders separated by an interleaver. Each constituent encoder can be any form of FEC coder used for conventional data communications. However, in practice, the constituent encoders are identical convolutional

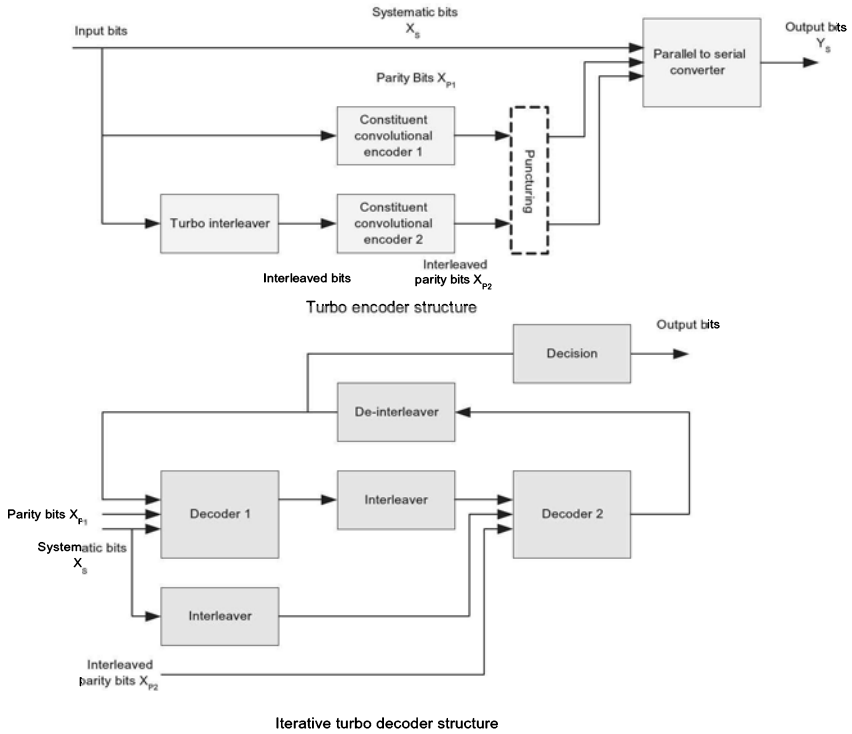


FIGURE 9.69

Generic structure of turbo encoder and iterative decoder [27,40].

encoders. As can be seen in Figure 9.69, the turbo coder consists of two identical constituent encoders. The input data stream and the parity outputs of the two parallel encoders are then serialized into a single turbo codeword. The interleaver is a critical component of the turbo code. It is a simple functional block that rearranges the order of the data bits in a prescribed, but irregular, manner. Although the same set of data bits is present at the output of the interleaver, the order of these bits has been changed. It must be noted that the interleaver used by a turbo coder is quite different than the rectangular interleavers that are commonly used in wireless systems to help mitigate the effect of deep fading. While a rectangular channel interleaver tries to space the data out according to a regular pattern, a turbo interleaver randomizes the ordering of the data in an irregular manner.

The operation of a turbo encoder and decoder can be described as follows. Let the binary input bits $\{0, 1\}$ be represented by bipolar levels $\{+1, -1\}$ and assigned to the variable d, which may take on the values $d = +1$ and $d = -1$. For an AWGN channel, the conditional probability density functions $f(x|d = -1)$

Samsung Ex. 1005, Page 560 of 1147

and $f(x|d=+1)$ are referred to as likelihood functions. The common hard-decision criterion, known as ML, selects the symbol $d_k=+1$ or $d_k=-1$ depending on the intercept point of received signal value x_k and the above conditional probability density functions using a fixed threshold of λ (decision point); thus, $d_k=+1$ if $x_k>\lambda$, otherwise, $d_k=-1$. Another decision rule, known as a maximum *a posteriori* probability (MAP) criterion takes into account the *a posteriori* probabilities $f(d=+1|x)$ and $f(d=-1|x)$ to construct the hypotheses H_1 and H_2 as follows:

$$f(d=+1|x) \underset{H_2}{\overset{H_1}{\gtrless}} f(d=-1|x) \tag{9.43}$$

That is, one selects hypothesis $H_1(d=+1)$, if the $f(d=+1|x)>f(d=-1|x)$; otherwise, $H_2(d=-1)$ is selected. Using the Bayes' theorem [30], the above *a posteriori* probabilities can be replaced by their equivalent expressions, yielding:

$$f(x|d=+1)p(d=+1) \underset{H_2}{\overset{H_1}{\gtrless}} f(x|d=-1)p(d=-1) \tag{9.44}$$

The likelihood ratio test is constructed as follows:

$$\frac{f(x|d=+1)p(d=+1)}{f(x|d=-1)p(d=-1)} \underset{H_2}{\overset{H_1}{\gtrless}} 1 \tag{9.45}$$

If the input bits are independent and identically distributed random variables, the above equation is simplified as follows:

$$\frac{f(x|d=+1)}{f(x|d=-1)} \underset{H_2}{\overset{H_1}{\gtrless}} 1 \tag{9.46}$$

By taking the logarithm of the above likelihood ratio, a useful metric known as log-likelihood ratio (LLR) is obtained, which is a real-valued number representing the soft-decision output of the detector:

$$L(d|x)=\log\left(\frac{f(x|d=+1)}{f(x|d=-1)}\right)+\log\left(\frac{p(d=+1)}{p(d=-1)}\right)=L(x|d)+L(d) \tag{9.47}$$

where $L(x|d)$ is the LLR of test statistics x obtained through measuring the channel output x under the condition that either $d=+1$ or $d=-1$ may have been transmitted, and $L(d)$ denotes the *a priori* LLR of the data bit d as shown in Figure 9.70. The introduction of a decoder would improve the reliability of the above decision-making process. For a systematic code, it can be shown that the LLR (soft output) expression at the output of the decoder can be written as $L_{\text{output}}(\hat{d})=L_{\text{input}}(\hat{d})+L_e(\hat{d})$, where $L_{\text{input}}(\hat{d})$ is the LLR of a data bit at the output of the detector (or the input to the decoder), and $L_e(\hat{d})$ represents additional

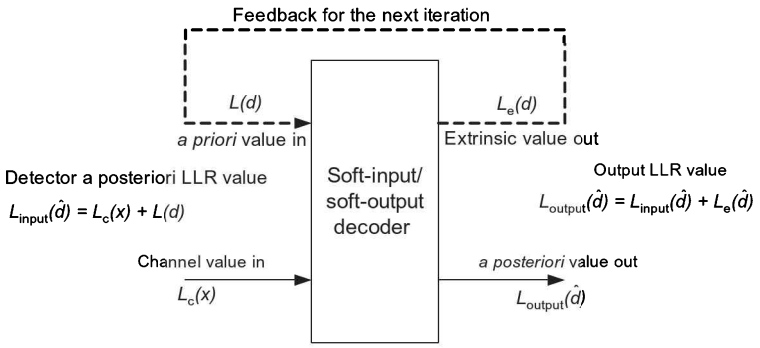


FIGURE 9.70

Illustration of soft-input/soft-output decoder for turbo codes [40,102].

information that is obtained from the decoding process (see Figure 9.70). The output sequence of a systematic decoder consists of values representing data and parity bits. Thus, the decoder output LLR can be decomposed into a data component that is associated with the detector measurement and the extrinsic component that is represented by the decoder contribution due to parity. The soft-decision component $L_{\text{output}}(\hat{d})$ is a real number that provides a hard-decision as well as the reliability of that decision. The sign of the parameter $L_{\text{output}}(\hat{d})$ denotes the hard-decision, i.e., for positive values of $L_{\text{output}}(\hat{d})$, decide $\hat{d} = +1$ and for negative values of $L_{\text{output}}(\hat{d})$, decide $\hat{d} = -1$. The magnitude of $L_{\text{output}}(\hat{d})$ denotes the reliability of the decision [98,102].

Turbo decoding relies on the exchange of probabilistic information between the two soft-input soft-output decoders shown in Figure 9.69. The extrinsic information is the result of the decoder's estimation of bit d, but not taking its own input into account. It is precisely this extrinsic information that is exchanged iteratively between the two soft-decision decoders during the decoding process. Subtracting the decoder's input from its output prevents the decoder from acting as a positive feedback amplifier and introduces stability in the feedback process. Usually, after a given number of iterations, one observes that the two decoders converge to a stable final decision for d. In practice, depending on the nature of the soft-decision decoder, scaling or clipping operations may be applied to the extrinsic information in order to ensure convergence within a small number of iterations. It is shown in the literature that the Max-Log-MAP decoding algorithm can make a good trade-off between performance and complexity, with the added advantage of not requiring any knowledge of the noise level (Figure 9.71). A stop criterion facilitates the convergence of the iterative decoding process and helps reduce the average power consumption of the decoder by reducing the average number of iterations required to decode a block without compromising performance [100–108].

Samsung Ex. 1005, Page 562 of 1147

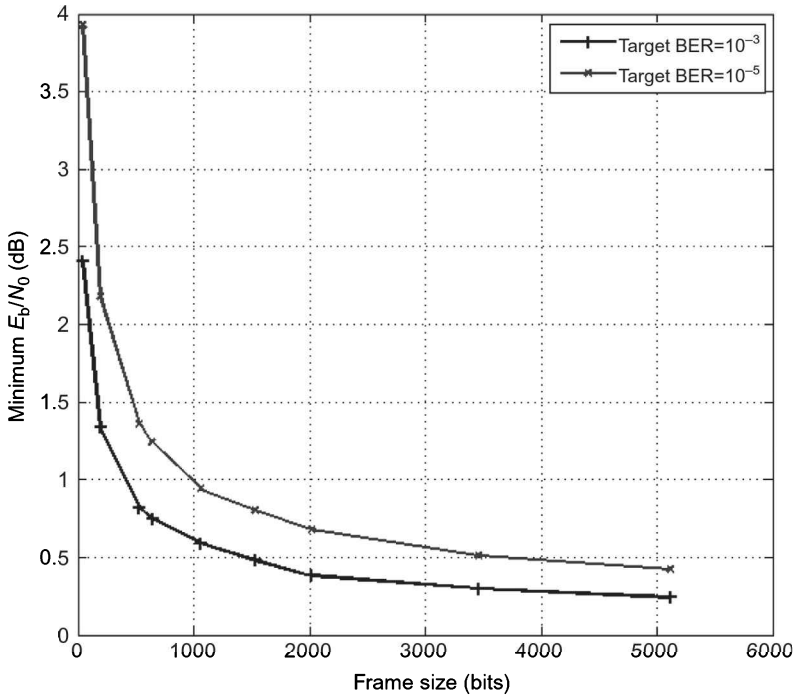


FIGURE 9.71

Minimum $E_b = N_0$ to achieve a target BER using the UMTS/LTE turbo coder ($R = 1/3$) with BPSK modulation over an AWGN channel [144].

For practical applications, it is often necessary to adapt the turbo code parameters to those of the application. The systematic bits are always present at the output of the turbo encoder; however, depending on the desired code rate, the parity bits from the two constituent encoders may be punctured prior to transmission (see Figure 9.69). The puncturing mechanism involves deleting some of the parity bits prior to transmission. As an example, if the minimum code rate of a turbo coder is 1/5, a code rate of 1/3 can be generated by deleting the second parity output of each encoder. Note that the puncturing pattern for each code rate must be known by the decoder for correct interpretation of the received bits [40].

It is shown that for smaller input frame sizes, less iteration is required; however, the performance of the turbo coder is improved by increasing the frame size. This is due to an increase in interleaver gain as the input frame size becomes larger. Figure 9.71 shows the minimum E_b/N_0 required to achieve a BER of 10^{-3} and 10^{-5} for various frame sizes using a UMTS/LTE turbo encoder with a minimum code rate of $R = 1/3$. In general, turbo codes are more suited for applications with large packet sizes and their performance degrades for small packet sizes.

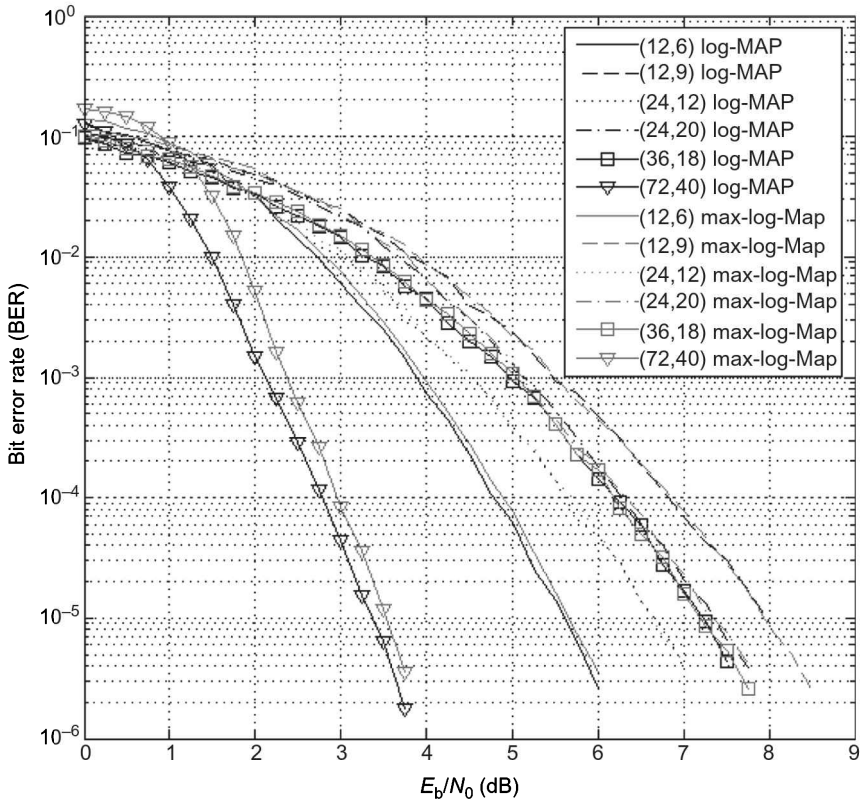


FIGURE 9.72

Comparison of the BER performance of a BTC turbo coder in AWGN channel with different code rates, frame sizes, and detector types [144].

Furthermore, the delay due to turbo interleaver might become prohibitive for delay-sensitive applications. Figure 9.72 shows the BER performance of a BTC turbo coder as a function of E_b/N_0 for various code rates, frame sizes, and detector types. In Figure 9.72, notation (n, k) denotes the number of input k and output n bytes [144].

9.11.3 Baseband digital modulation schemes

Modulation is a process by which a carrier signal is altered according to information in a message signal. For the typical baseband modulation types (PSK, DPSK, FSK, QAM, or PAM), gray-coded constellations are obtained by selecting the gray parameter in the corresponding baseband modulation function. The mapping

or assignment of k information bits to $M = 2^k$ possible signal amplitudes may be done in a number of ways. The preferred assignment is the one in which the adjacent signal amplitudes differ by one binary digit since the most likely errors caused by noise involve the erroneous selection of an adjacent amplitude to the transmitted signal amplitude. In that case, only a single bit error occurs in the k-bit sequence. This mapping is referred to as gray coding. The criteria that must be taken into account when choosing a digital modulation method include the following:

- Power efficiency, i.e., the E_b/N_0 ratio for a specific error probability
- Bandwidth efficiency, i.e., the data rate per unit bandwidth
- Performance in multipath fading channels and under non-linear distortion
- Implementation cost and complexity

However, the above requirements cannot be satisfied at the same time, and some trade-off is necessary. In order to explain the latter statement, let us define a few basic terms in this context. A constellation diagram is a graphical representation of the complex envelope of each possible symbol state. The power efficiency is related to the minimum distance between the points in the constellation. The bandwidth efficiency is related to the number of points in the constellation. The gray coding is used to assign groups of bits to each constellation point. In gray coding, adjacent constellation points differ by a single bit.

The two most common digital modulation schemes used in cellular systems are M-ary PSK and M-ary QAM modulation schemes. In M-ary PSK, a group of n bits are transmitted using $M = 2^n$ different phases, whereas M-ary QAM uses a combination of amplitude and phase modulation to convey the information. Study of the literature suggests that the minimum code distance d_{min} is smaller for M-ary ASK than for M-ary PSK or M-ary QAM for the same average transmitted power, which translates into poorer error performance because the signal envelope is not constant, it results in poor performance under non-linear distortion condition. The minimum distance d_{min} is smaller for M-ary PSK compared to that for M-ary QAM for the same average transmitted power, resulting in poorer error performance. However, a constant envelope results in good performance under nonlinear distortion. The minimum distance d_{min} is larger for M-ary QAM relative to that for M-ary ASK or M-ary PSK for the same average transmitted power, which means better error performance. Nonetheless, since the signal envelope is not constant, poor error performance is expected under non-linear distortion. Pulse shaping is then used to reduce the bandwidth requirements of the modulated signal. The Nyquist technique can be used to reduce the bandwidth requirement and to eliminate ISI. The non-Nyquist techniques reduce the bandwidth requirement, but do not eliminate ISI. Note that in all M-ary modulation schemes, a group of n bits is transmitted in each signaling interval $T = T_b \log_2 M$ and the modulated signal bandwidth is inversely proportional to the signaling interval. It can be shown that the probability of error (i.e., error per bit) equations as a function of E_b/N_0 for M-ary