Textbooks in Telecommunication Engineering

Wei Jiang
Bin Han

Cellular Communication Networks and Standards

The Evolution from 1G to 6G

Springer

Textbooks in Telecommunication Engineering

Series Editor

Tarek S. El-Bawab, Ph.D., Boston, Massachusetts, USA

Telecommunication and networks have evolved to embrace all aspects of our everyday life, including smart cities and infrastructures, healthcare, banking and businesses, manufacturing, space and aviation, meteorology and climate change, oceans and marine life, Internet of Things, defense, homeland security, education, research, social media, entertainment, and many others. Network applications and services continue to expand, virtually without limits. Therefore, specialized telecommunication and network engineering programs are recognized as a necessity to accelerate the pace of advancement in this field, and to prepare a new generation of engineers for imminent needs in our modern life. These programs need curricula, courses, labs, and textbooks of their own.

The IEEE Communications Society's Telecommunication Engineering Education (TEE) movement, led by Tarek S. El-Bawab- the editor of this Series, resulted in recognition of this field of engineering by the Accreditation Board for Engineering and Technology (ABET) on November 1, 2014. This Springer Series was launched to capitalizes on this milestone. The Series goal is to produce high-quality textbooks to fulfill the education needs of telecommunication and network engineering, and to support the development of specialized undergraduate and graduate curricula in this regard. The Series also supports research in this field and helps prepare its scholars for global challenges that lay ahead. The Series have published innovative textbooks in areas of network science and engineering where textbooks have been rare. It is producing high-quality volumes featuring innovative presentation media, interactive content, and online resources for students and professors.

Book proposals are solicited in all topics of telecommunication and network engineering including, but not limited to: network architecture and protocols; traffic engineering; network design, dimensioning, modeling, measurements, and analytics; network management and softwarization; cybersecurity; synchronization and control; applications of artificial intelligence in telecommunications and networks; applications of data science in telecommunications and networks; network availability, reliability, protection, recovery and restoration; wireless communication systems; cellular technologies and networks (through 5G, 6G, and beyond); satellite and space communications and networks; optical communications and networks; heterogeneous networks; broadband access and free-space optical communications; MSO/cable networks; storage networks; optical interconnects; and data centers; social networks; transmission media and systems; switching and routing (from legacy to today's paradigms); network applications and services; telecom economics and business; telecom regulation and policies; standards and standardization; and laboratories.

Proposals of interest shall be for textbooks that can be used to develop university courses, either in full or in part. They should include recent advances in the field while capturing whatever fundamentals that are necessary for students to understand the bases of the topic and appreciate its evolution trends. Books in this series will provide high-quality illustrations, examples, end-of-chapters' problems/exercises and case studies.

For further information and to submit proposals, please contact the Series Editor, Dr. Tarek S. El-Bawab, telbawab@ieee.org; or Mary James, Executive Editor at Springer, mary.james@springer.com.

This series is indexed in Scopus.

Wei Jiang • Bin Han

Cellular Communication Networks and Standards

The Evolution from 1G to 6G

Springer

Wei Jiang
German Research Center for
Artificial Intelligence (DFKI)
Kaiserslautern, Germany

Bin Han
University of Kaiserslautern-Landau (RPTU)
Kaiserslautern, Germany

ISSN 2524-4345 ISSN 2524-4353 (electronic)
Textbooks in Telecommunication Engineering
ISBN 978-3-031-57819-9 ISBN 978-3-031-57820-5 (eBook)
https://doi.org/10.1007/978-3-031-57820-5

Preface

As the commercial deployment of fifth-generation (5G) cellular communication networks is underway worldwide, attention from both academia and the telecommunication industry is already shifting toward the development of the sixth-generation (6G) technologies. The significant impact of 5G has raised awareness among governments and the public regarding the vital role of mobile systems in economic prosperity and national security. Over the past few years, numerous countries and major tech companies have unveiled ambitious plans and initiated research endeavors for 6G development. A pivotal moment occurred in 2023 when the International Telecommunication Union (ITU) Radiocommunication Assembly (RA-23) officially designated the next generation of International Mobile Telecommunications (IMT) as *IMT-2030*. Another milestone was achieved in November 2023 when the ITU Radiocommunication Sector (ITU-R) Working Party 5D (WP 5D) published a new recommendation for the IMT-2030 framework, known as ITU-R *M.2160: Framework and Overall Objectives of the Future Development of IMT for 2030 and Beyond*. This framework specified six usage scenarios for IMT-2030, encompassing Immersive Communication, Hyper-Reliable and Low-Latency Communication, Massive Communication, Ubiquitous Connectivity, Artificial Intelligence and Communication, and Integrated Sensing and Communication. Additionally, the World Radiocommunication Conference 2023 (WRC-23) identified the spectrum for IMT toward 2030, a critical step for expanding broadband connectivity and advancing IMT mobile services in the forthcoming era of 6G.

Since Guglielmo Marconi's successful demonstration of radio transmission in the summer of 1895, we have witnessed a remarkable journey spanning over a century, culminating in the evolution of state-of-the-art mobile communications. Throughout the years, a wide range of radio communications and broadcasting services has been embraced on a global scale. In the pivotal year of 1947, Bell Labs, a globally renowned research institution, achieved a groundbreaking innovation—the Cellular Network. Fast forward to the early 1980s, after extensive technical development, the introduction of first-generation cellular networks marked a significant milestone, offering commercial mobile telephony services to the public. The advantages of ease of deployment, economic efficiency, portability, flexibility, and scalability inherent in mobile cellular networks, as opposed to wire-line networks, fueled explosive growth in the ensuing decades. These networks emerged as critical infrastructures, playing a pivotal role in empowering modern society and reshaping human behaviors across various domains such as business, education, entertainment, and personal life.

To comprehend the complexities of cutting-edge cellular technologies and systems, a comprehensive understanding of their evolution is essential. The motivation behind this textbook is to offer students and wireless engineers a thorough review of the entire history of mobile cellular systems, from pre-cellular to the sixth generation. By doing so, readers can gain insights into the upcoming 6G and beyond systems. At this historical juncture of developing 6G, it is strongly believed that this book will serve as an enlightening guide, igniting interest and prompting further investigations into 6G and beyond communication systems. The book aims to attract a broad audience in academia and industry across all related fields.

Organized into 12 chapters, the textbook provides a structured exploration of the subject matter.

Chapter 1: Standards History of Cellular Communications Systems

This chapter provides a brief look at the history of cellular communication networks and standards, charting their evolution from initial analog communication systems toward the upcoming 6G intelligent networks. By offering this thorough examination, readers will develop an initial insight into the pivotal moments that have defined the journey of cellular communications and the breakthroughs poised to influence its trajectory moving forward.

Chapter 2: Evolution to First-Generation (1G) Mobile Cellular Communications

In this chapter, we offer a thorough examination of 1G systems, based on a revolutionary innovation of an elegant network design known as the cellular system. Initially, we present a succinct summary of pre-cellular systems to contextualize the nascent stages of mobile communications. The chapter subsequently compares the diverse standards constituting 1G, originating from different countries. Finally, we explore the key technologies that propelled the evolution of 1G, aiming for a comprehensive understanding of the historical and technical foundations underlying this early phase in cellular communication networks and standards.

Chapter 3: Evolution to Second-Generation (2G) Mobile Cellular Communications

The success of 1G cellular communication networks and standards brought about a revolutionary shape in the telecommunication industry. However, its rudimentary design and technological constraints proved inadequate to meet the escalating market demands. The progress in digital technology during the 1980s facilitated the transition from 1G analog cellular to 2G digital cellular communication networks and standards, presenting numerous benefits such as increased capacity, enhanced service quality, and improved security. This chapter explores the key factors driving the shift from 1G to 2G, introducing diverse standards within the realm of 2G.

Chapter 4: The Global System for Mobile Communications (GSM)

This chapter undertakes an in-depth exploration of GSM, the most successful 2G cellular communication standard that significantly transformed the telecommunications industry. The objective is to furnish readers with a comprehensive grasp of GSM's architecture, its essential technological components, and the challenges encountered throughout its evolution. The various elements constituting the GSM network are scrutinized, with their functionalities examined in detail. Additionally, the chapter presents the protocols and interfaces facilitating seamless communication within the GSM network.

Chapter 5: Evolution to Third-Generation (3G) Mobile Cellular Communications

This chapter clarifies the big leap in cellular communication networks and standards from being voice-centric to data-centric, adapting to new user behaviors and unprecedented traffic patterns of Internet-based data services. This chapter clarifies the major driving forces behind the transition from 2G to 3G, and summarizes the main technical standards that constitute 3G, including WCDMA, CDMA2000, TD-SCDMA, and WiMAX. Then, we study the fundamentals of code-division multiple access (CDMA) that propelled the evolution of 3G, along with other key 3G technologies.

Chapter 6: Universal Mobile Telecommunications Service (UMTS)

In this chapter, we focus on UMTS, the mainstream 3G standard that played a critical role in the advancement of cellular communication networks and standards. This chapter provides an in-depth analysis of UMTS's system architecture, its interfaces, the design considerations, and the unique features that set UMTS apart from the previous generations, such as advanced roaming capabilities and support for location-based services. The chapter also examines the challenges and limitations of UMTS, offering insights into how it fits into the broader context of mobile communications evolution.

Chapter 7: Evolution to Fourth-Generation (4G) Mobile Cellular Communications

This chapter highlights the first unified cellular communication standard worldwide in the history of the telecommunications industry. The aim is to review 4G cellular communication networks and standards and shed light on the driving forces behind the transition from 3G data-centric cellular networks to 4G mobile broadband. The fundamental technologies empowering

the success of 4G, such as MIMO, OFDM, OFDMA, SC-FDMA, relaying, and D2D, are elaborated to provide readers with an insightful view.

Chapter 8: Long-Term Evolution Advanced (LTE-A)

This chapter explores the 4G's successor of UMTS: LTE-A, based on 3GPP Release 10. We give a comprehensive overview of LTE-A, focusing on its key features, system architecture, radio interface protocols, mobility management, and security. The significant advancements and differences from its predecessor, UMTS, are highlighted. Concentrating on Release 10, the first standard meeting the 4G requirements proposed in IMT-Advanced, this chapter does not cover technologies introduced in later releases of LTE-A.

Chapter 9: Evolution to Fifth-Generation (5G) Mobile Cellular Communications

The ninth chapter clarifies the most significant feature of 5G, also known as IMT-2020, which goes beyond connecting people to interconnecting humans, machines, and things. This chapter provides readers with a comprehensive view of 5G, covering its driving forces, three usage scenarios (enhanced mobile broadband, ultra-reliable low-latency communications, and massive machine-type communications), key performance indicators, the standardization process, and key technological enablers.

Chapter 10: New Radio (NR) and 5G Core Networks

In this chapter, we study the main features in 3GPP Release 15, symbolizing the shift from 4G to 5G technologies. The evolution of the Radio Access Network (RAN) places a spotlight on innovations within the NR framework. The chapter underscores the significance of the 5G Core (5GC) with its service-based architecture and examines the role of network slicing in optimizing resource utilization. Also, a comprehensive examination of the security enhancements in 5G, emphasizing the need to protect the network from diverse threats, is offered.

Chapter 11: Evolution Toward Sixth-Generation (6G) Mobile Cellular Systems

This chapter seeks to offer readers insights into the current state and future trajectory of evolving cellular communication networks and standards, specifically focusing on the progression toward 6G and beyond. We explore the driving forces behind 6G development, project significant traffic growth by 2030, envision potential use cases and applications, outline the six IMT-2030 usage scenarios recently defined by ITU-R, estimate performance capacities, summarize global research initiatives, and provide a glimpse into the anticipated roadmap for the deployment of 6G in 2030.

Chapter 12: Key Technologies for Sixth-Generation (6G) Mobile Cellular Systems

The last chapter of this book discusses potential key technologies for achieving ultra-high performance in disruptive 6G use cases and applications. Organized into categories, it covers *New Spectrum* possibilities in terahertz (THz) and optical bands; advancements in the *New Air Interface*, including cell-free massive MIMO, ultra-massive MIMO, intelligent reflecting surfaces, and next-generation multiple access; prospects in *New Networking*, encompassing open radio access networks (O-RAN) and non-terrestrial networks; and the profound impact of the *New Paradigm*, driven by the convergence of communication, artificial intelligence (AI), and sensing.

For Whom Is This Book Written?

This book aims to transcend being merely a comprehensive reference for professionals involved in the development, standardization, deployment, and applications of 6G systems. While it caters to a wide audience including engineers, researchers, manufacturers, network operators, software developers, content providers, service providers, broadcasters, and regulatory bodies, its primary ambition is to serve as an enriching textbook for graduate students in diverse fields such as circuits design, signal processing, electronic engineering, wireless communications, artificial intelligence, microwave technology, information theory, antenna

and propagation, system-on-chip implementation, and computer networks, offering a holistic educational resource for the upcoming generation of experts in cellular communications industry.

Kaiserslautern, Germany
February 27, 2024

Wei Jiang
Bin Han

Contents

Acronyms

16-QAM	16-ary quadrature amplitude modulation. 32
1G	First Generation. 1, 2, 6, 21, 23, 37, 48, 49, 168, 231
2D	two-dimensional. 200, 201
2G	Second Generation. 1–3, 6, 21, 53, 69, 71, 72, 231, 234
32-QAM	32-ary quadrature amplitude modulation. 32
3D	three-dimensional. 158, 200
3G	Third Generation. 1–6, 31, 32, 50, 53, 54, 69, 104, 107, 123, 143, 231, 233, 234
3GPP	*3rd Generation Partnership Project*. 1, 3–6, 26, 42, 55, 61, 63, 69, 72, 74, 75, 77, 80, 86–88, 107, 108, 125, 126, 130, 133, 145–147, 158, 159, 168, 169, 173, 174, 181, 183–185, 188–192, 231, 234, 235
3GPP2	*3rd Generation Partnership Project 2*. 1, 3, 4, 6, 55, 107, 231
4G	Fourth Generation. 1, 2, 4–6, 107, 115, 123, 125, 134, 147, 149, 168, 169, 171, 183, 189, 231, 234–236
5G	Fifth Generation. 4–6, 149, 157–160, 168–181, 183–193, 195, 196, 231, 236, 237
5GC	5G Core. 5, 156, 169, 171, 172, 179, 183–186, 188, 190, 192, 193, 236
5GS	5G System. 156, 157, 169, 186, 188
6G	Sixth Generation. 5, 6, 195, 196
8PSK	eight phase-shift keying. 31, 32, 58
A-GNSS	Assisted Global Navigation Satellite System. 101
ACB	Access Class Barring. 188
ACK	Acknowledgements. 62, 63, 89, 127, 137
ADC	administration center. 39
AF	Application Function. 185, 186, 190, 192
AGCH	Access Grant Channel. 41
AI	artificial intelligence. vii, 197, 200, 201, 212, 224, 230, 231, 237, 238
AICH	Acquisition Indicator Channel. 75, 88, 89
AKA	authentication and key agreement. 96–98, 146, 147, 185, 236
AM	acknowledged mode. 91, 94
AM	amplitude modulation. 13, 97
AMBR	aggregate MBR. 138, 139
AMC	adaptive modulation and coding. 62, 65, 68
AMF	Access and Mobility Management Function. 185, 186, 188, 190, 192, 193, 237
AMPS	Advanced Mobile Phone System. 1, 2, 12, 24
ANSI	*American National Standards Institute*. 3
AoSA	array-of-subarrays. 217
AP	access point. 218
API	Application Programming Interface. 157
APN	Access Point Name. 138, 139

AR augmented reality. 195, 199, 201
ARIB *Association of Radio Industries and Businesses*. 3, 55–57
ARP allocation retention priority. 138, 139
ARPF Authentication Credential Repository and Processing Function. 190
ARQ automatic repeat request. 3, 68, 75, 111, 137, 181
AS Application Servers. 103
ASC access service class. 88, 89, 91
ATIS *Alliance for Telecommunications Industry Solutions*. 55, 56, 196
AuC Authentication Center. 30, 39, 45, 46, 72, 96–98, 234
AUSF Authentication Server Function. 185, 186, 190, 192, 193, 237

BBU Baseband Processing Unit. 180
BCC base transceiver station color code. 41
BCCH Broadcast Control Channel. 41, 42, 72, 73
BCFE Broadcast Control Functional Entity. 93
BCH Broadcast Channel. 41, 73–75, 80, 83, 94, 95, 174
BER bit error rate. 68
Bm mobile B. 41
BMC broadcast/multicast control. 90, 93, 136, 235
BPSK binary phase shift keying. 58, 77
BS base station. 38, 119
BSC Base Station Controller. 30, 38, 72, 129, 234
BSIC Base Transceiver Station Identity Code. 41
BSS Base Station Subsystem. 30, 37–39, 46, 47, 233
BTS Base Transceiver Station. 13, 14, 30, 38, 41, 42, 46, 50, 72, 234

C-Netz Radio Telephone Network C. 1, 2
C-RAN Centralized RAN. 179, 180
C-SAP Control SAP. 92
CA carrier aggregation. 108, 111, 126, 131, 133, 134, 172
CAMEL Customized Applications for Mobile networks Enhanced Logic. 102
CAPEX capital expenditure. 122, 123, 151, 224
CATT Chinese Academy of Telecommunications Technology. 59
CB cell broadcast. 93
CBC Cell Broadcast Center. 93
CBCF Cell Broadcast Centre Function. 186
CBS cell broadcast service. 93, 96
CC country code. 40, 41
CC component carrier. 126
CCCH Common Control Channel. 41, 72, 73, 99
CCPCH Common Control Physical Channel. 127
CCSA *China Communications Standards Association*. 55–57
CDD cyclic delay diversity. 112
CDMA Code-Division Multiple Access. 2–4, 20, 24, 55, 56, 65, 69, 71, 84, 115, 137, 162, 168, 223
CDMA code-division multiple access. 231, 234
CDMA2000 Code-Division Multiple Access 2000. 31, 55, 61, 64, 69
CEPT *European Conference of Postal and Telecommunications Administrations*. 2, 25, 37, 231
CHF Charging Function. 185, 186
CI cell identifier. 41
CM connection management. 93
CN core network. 57, 72, 97, 102, 128, 171, 188, 234

CoMP	coordinated multi-point transmission and reception. 110, 111, 119, 125, 147, 180, 218, 235, 236
CP	control plane. 90, 173, 174, 177, 185, 190, 191
CP-OFDM	cyclic prefix OFDM. 174, 236
CPC	continuous packet connectivity. 63, 64
CPICH	Common Pilot Channel. 75, 87, 88
CPRI	Common Public Radio Interface. 181
CQI	channel quality information. 62, 87, 127, 137, 138
CRC	cyclic redundancy check. 34, 36, 68, 80, 84, 86, 87
CRS	Cell Specific RS. 175
CS	circuit switched. 97, 146
CSCF	Call Session Control Function. 103, 106, 235
CSI	channel state information. 113, 119, 131, 138, 161, 162, 175, 218, 227
CSI-RS	channel state information reference signal. 131, 161, 175
CTCH	Common Traffic Channel. 72, 73, 93
CU	Centralized Unit. 179–181, 190, 191, 193, 237
CWTS	China Wireless Telecommunication Standards. 55
D-AMPS	Digital Advanced Mobile Phone System. 14, 24, 26
D2D	device-to-device. 110, 111, 121–123, 125, 147, 213
DC	Dual connectivity. 181
DC-HSDPA	dual-carrier HSDPA. 64
DC-SAP	Dedicated Control SAP. 94
DCCH	Dedicated Control Channel. 41, 72, 73, 99, 127
DCFE	Dedicated Control Functional Entity. 93, 94
DCH	Dedicated Channel. 73–75, 80, 83, 94, 127
DCI	Downlink Control Information. 127, 174
DCS-1800	Digital Cellular System 1800MHz. 26, 232
DECT	Digital Enhanced Cordless Telecommunications. 56, 61
DenB	Donor eNodeB. 134
DFT	discrete Fourier transform. 114, 235
DFT-s-OFDM	discrete Fourier transform spread OFDM. 116, 174
DHCP	Dynamic Host Configuration Protocol. 185
DL	downlink. 3, 4, 38, 41, 42, 51, 71–75, 77, 80–85, 87–89, 96, 97, 126, 127, 137, 138, 141, 143, 145, 146, 158, 174–176, 178, 233
DLC	data link control. 60, 61
DM-RS	demolulation RS. 175
DN	Data Network. 184, 186
DPCCH	Dedicated Physical Control Channel. 62, 74, 75, 78, 87, 88
DPCH	Dedicated Physical Channel. 75, 77, 79, 94
DPDCH	Dedicated Physical Data Channel. 74, 75, 78, 79, 87, 88
DRX	discontinuous reception. 138
DS-CDMA	direct-sequence CDMA. 56
DSS	dynamic spectrum sharing. 177
DSSS	direct-sequence spread spectrum. 65
DTCH	Dedicated Traffic Channel. 72, 73, 127
DTX	discontinuous transmission. 35, 82–84, 87, 138
DU	Distributed Unit. 179–181, 190, 191, 193, 237
E-AGCH	E-DCH Absolute Grant Channel. 63, 74, 75, 88
E-DCH	Enhanced Dedicated Channel. 63, 73–75, 91, 127
E-DPCCH	E-DCH Dedicated Physical Control Channel. 63, 74, 75, 79, 88
E-DPDCH	E-DCH Dedicated Physical Data Channel. 63, 74, 75, 79, 88

E-HICH E-DCH Hybrid ARQ Indicator Channel. 63, 75, 88
E-RGCH E-DCH Relative Grant Channel. 63, 75, 88
E-UTRA Evolved Universal Terrestrial Radio Access. 107, 125, 126, 131, 133, 172–175, 177, 182
E-UTRAN Evolved Universal Terrestrial Radio Access Network. 125, 128, 129, 131, 137, 141, 142, 145–147, 169, 175–177, 179, 188, 235
E2E end-to-end. 158, 159, 169, 172, 181, 200, 202
EC European Commission. 196
ECSD Enhanced Circuit Switched Data. 31
EDGE Enhanced Data Rates for GSM Evolution. 1–3, 24, 26, 29, 31, 42, 53, 60, 61, 71, 72
EGPRS Enhanced General Packet Radio Service. 3, 31
EIA *Electronic Industries Alliance*. 3
eICIC enhanced ICIC. 110, 119, 120, 133
EIR Equipment Identity Register. 30, 39, 72, 185, 186
eMBB enhanced mobile broadband. 5, 119, 152, 156, 167, 175, 182, 183, 198, 200, 202, 205, 212, 223, 231, 236
eMBMS enhanced Multimedia Broadcast Multicast Services. 110
EN-DC E-UTRA-NR Dual Connectivity. 177, 188
eNodeB E-UTRAN Node B, or eNB. 129, 130, 133, 134, 136–141, 143, 179, 235, 236
EPC Evolved Packet Core. 4, 107–109, 128–130, 169, 171, 183, 185, 188, 235, 236
EPS Evolved Packet System. 138, 139, 146, 147, 169, 186, 188, 236
ETSI *European Telecommunications Standards Institute*. 25, 26, 43, 44, 55, 56, 61
EU *European Union*. 54, 237
EV-DO Evolution-Data Optimized. 2–4
EV-DV EVolution-Data Voice. 2, 4

F-DPCH Fractional Dedicated Physical Channel. 75, 87, 88
FAC final assembly code. 39
FACCH Fast Associated Control Channel. 41
FACH Forward Access Channel. 73–75, 80, 83, 94, 127
FC fully-connected. 217
FCC *Federal Communications Commission*. 12, 13, 196, 226
FCCH Frequency Correction Channel. 41, 42
FDD frequency-division duplex. 4, 11–13, 15–17, 20, 21, 42, 55, 56, 71, 72, 87, 88, 108, 109, 125, 131, 133, 143, 161, 174, 231, 233, 234
FDMA freuqnecy-division multiple access. 2, 12, 13, 15–17, 33, 42, 114–116, 162, 168, 231, 233, 234
FEC forward error correction. 68, 80
FFR fractional frequency reuse. 133
FFSK fast frequency-shift keying. 15
FFT fast Fourier transform. 64, 173
FH frequency hopping. 42, 43
FM frequency modulation. 1, 9, 12, 13, 15, 16
FOMA Freedom of Mobile Multimedia Access. 54, 56
FPLMTS Future Public Land Mobile Telecommunications System. 53, 203
FR frequency range. 172
FSK frequency-shift keying. 1, 13, 16, 41
FSTD frequency-switched transmit diversity. 112

GBR guaranteed bit rate. 138, 139

GC-SAP General Control SAP. 94
GCI global cell identity. 41
GEO geostationary Earth orbit. 202, 225, 238
GERAN GSM EDGE Radio Access Network. 26, 57, 72, 128, 129, 141, 145, 234
GGSN Gateway GPRS Support Node. 30, 72, 234
GMLC Gateway Mobile Location Centre. 102, 104
GMSC gateway MSC. 39, 72
GMSK Gaussian Minimum Shift Keying. 31, 34
GMSK Gaussian MSK. 41, 233
gNodeB Next-Generation Node B, or gNB. 179, 190, 191, 237
GPRS General Packet Radio Service. 2–4, 24, 26, 29, 31, 53, 61, 71, 72, 97, 129, 231
GSM Global System for Mobile Communications. 2–4, 15, 24, 26, 35, 37–51, 53, 55, 61, 69, 71, 72, 96, 97, 99, 101, 102, 106, 129, 231, 233, 234

HARQ hybrid automatic repeat request. 31, 58, 59, 61–65, 68, 75, 84–87, 91, 127, 136–138, 163, 176, 181, 235
HC header compression. 92
HCS hierarchical cell structure. 4
HE Home Environment. 97, 98
HetNet Heterogeneous Network. 134, 236
HLR Home Location Register. 30, 39–41, 46–48, 72, 96, 97, 234
HRPD High-Rate Packet Data. 58
HS-DPCCH Dedicated Physical Control Channel (uplink) for HS-DSCH. 74, 75, 78, 79, 87, 88
HS-DSCH High Speed Downlink Shared Channel. 62, 73–75, 84–86, 91, 127
HS-PDSCH High Speed Physical Downlink Shared Channel. 74, 75, 77, 86, 88
HS-SCCH Shared Control Channel for HS-DSCH. 62, 74, 75, 86, 88
HSCSD High Speed Circuit Switched Data. 2, 3, 24, 29, 31
HSDPA High Speed Downlink Packet Access. 2, 3, 61–64, 108, 136, 138
HSPA High Speed Packet Access. 3, 60, 61, 134, 146
HSPA+ Evolved High Speed Packet Access. 2–4, 61, 63, 64
HSS Home Subscriber Server. 72, 102, 103, 129, 130
HSUPA High Speed Uplink Packet Access. 3, 58, 61, 63, 108, 136, 146
HTC human-type communication. 145
HTML HyperText Markup Language. 28
HTTP Hypertext Transfer Protocol. 29

I-CSCF Interrogating-CSCF. 103
ICIC inter-cell interference coordination. 119, 120, 133
ID identification. 37, 134
IEEE *Institute of Electrical and Electronics Engineers*. 1–4, 6, 59, 60, 107, 231
IMEI International Mobile Equipment Identity. 30, 37, 39, 40, 50
IMEI international mobile equipment identity. 233
IMS IP Multimedia Subsystem. 69, 72, 102–105, 139, 234, 235
IMSI International Mobile Subscriber Identity. 30, 39, 40, 45–47, 50, 98
IMSI international mobile subscriber identity. 233
IMT International Mobile Telecommunications. 1, 3–5, 125, 197–199, 203, 234, 235
IMT-2000 International Mobile Telecommunications-2000. 53–56, 60, 69, 107, 123
IMT-2020 International Mobile Telecommunications-2020. 150, 152, 154, 155, 168, 169, 200, 203, 236
IMT-2030 International Mobile Telecommunications-2030. 5, 195, 199, 200, 237

IMT-Advanced International Mobile Telecommunications-Advanced. 107–109, 123, 134, 147, 235
IoT Internet of Things. 6, 109, 125, 145, 146, 149–151, 157, 162, 189, 195, 198, 201–205, 212, 215, 222, 225, 229
IP Internet Protocol. 4, 53, 56, 69, 72, 102–104, 107, 108, 125, 130, 138, 139, 185, 231, 235
IPsec IP security. 185
IRC interference rejection combining. 134
IRS intelligent reflecting surface. 220–222, 238
ISAC integrated sensing and communication. 213, 226
ISDN Integrated Services Digital Network. 25, 29, 40, 41, 56
ISI inter-symbol interference. 33, 41, 65, 114
ITU *International Telecommunication Union*. 54–56, 72, 125, 234
ITU-R *International Telecommunication Union, Radiocommunication Sector*. 1, 3, 5, 53, 55, 56, 60, 69, 107, 108, 123, 150, 152, 196, 199, 203, 208, 211, 234, 238
ITU-T *International Telecommunication Union, Telecommunication Standardization Sector*. 195

Kc Ciphering Key. 45, 46
Ki Subscriber Authentication Key. 45–47
KPI key performance indicator. 5, 205

L-EBI Linked EPS bearer ID. 139
LA location area. 40, 41, 47, 48, 146, 233
LAA license-assisted access. 110, 111, 122, 125, 147
LAC LA code. 41
LAI LA identity. 40, 41, 46–48
LAPDm Link Access Protocol on the Dm Channel. 38
LBS location-based services. 71
LBT Listen-Before-Talk. 179
LCS location services. 101–104, 106, 234
LDPC low-density parity-check. 56, 60, 167, 172, 174, 222, 234, 236
LED light-emitting diode. 215
LEO low Earth orbit. 196, 202, 225, 226
Lm lower-rate mobile. 41
LMF Location Management Function. 185
LMS Least Mean Square. 134
LMSI local mobile subscriber identity. 40, 50, 233
LoS line-of-sight. 59, 217
LPC linear predictive coder. 43
LTE Long-Term Evolution. 1, 2, 4, 60, 63, 107, 108, 111, 114, 123, 125–128, 130–133, 136–138, 140–146, 169, 171, 174, 175, 177, 181, 186, 188, 192, 231, 235–237
LTE-A Long-Term Evolution Advanced. 1, 2, 4, 107, 125–143, 145–147, 169, 174, 175, 177, 186, 188, 193, 235–237
LTE-Advanced Long-Term Evolution Advanced. 107, 108, 114, 235
LU location update. 39, 40, 46, 47, 141, 233

M2M machine-to-machine. 151, 198, 199
MAC medium access control. 31, 60, 61, 80, 89, 90, 93, 134, 136, 137, 143, 147, 227, 235
MAP Mobile Application Part. 102

MBMS	multimedia broadcast/multicast service. 75, 91, 127
MBR	maximum bit rate. 138, 139
MC-CDMA	multi-carrier code-division multiple access. 116
MCC	mobile country code. 39
MCG	Master Cell Group. 177
MCH	Multicast Channel. 127
MCM	multi-carrier modulation. 114
MCS	modulation and coding scheme. 31, 62, 134, 138
ME	mobile equipment. 37–40, 45, 72, 96, 97, 233
MEC	multi-access edge computing. 181, 189
MGCF	Media Gateway Control Function. 103, 106, 235
MGW	Media Gateway. 103, 106, 235
MICH	MBMS Indicator Channel. 75, 88
MICO	Mobile Initiated Connection Only. 188
MIMO	multiple-input multiple-output. 2–4, 6, 56, 64, 86, 87, 108, 110, 111, 123, 125, 131, 138, 146, 147, 157, 160, 174, 181, 216–218, 223, 231, 234, 235, 238
MM	mobility management. 47, 48, 93, 142, 186–188
MME	Mobility Management Entity. 129, 130, 146, 186, 235, 236
MMS	multimedia messaging service. 26
MMSE	Minimum Mean Square Error. 134
mMTC	massive machine-type communication. 5, 152, 156, 167, 175, 182, 183, 200, 201, 206, 231, 236
mmWave	millimeter wave. 6, 59, 149, 160, 162, 163, 172, 174, 217, 222
MNC	mobile network code. 40, 41
MNO	mobile network operator. 1, 44
MPS	Multimedia Priority Services. 139
MR	mixed reality. 195, 201
MRC	maximum-ratio combining. 66, 68
MS	mobile station. 37–40, 42, 45–51, 72, 233
MSC	Mobile Switching Center. 30, 38, 39, 41, 46–48, 66, 72, 102, 104, 234
MSIN	mobile subscriber identification number. 40
MSISDN	mobile subscriber ISDN number. 40, 50
MSK	minimum-shift keying. 41
MSRN	mobile station roaming number. 40, 47
MSS	Mobile Station Subsystem. 30
MT	mobile terminal. 37, 38
MTC	machine-type communication. 4, 110, 145, 146
MTCH	Multicast Traffic Channel. 127
MTSO	Mobile Telephone Switching Office. 13, 14
MU-MIMO	multi-user multiple-input multiple-output. 131, 161
MUST	multi-user superposed transmission. 223
N3IWF	Non-3GPP Interworking Function. 185
NACK	Negative Acknowledgements. 62, 63, 89, 127, 137
NAS	Non-Access Stratum. 136, 147, 185, 190, 192, 236
NB-IoT	narrow-band Internet of Things. 4, 110, 146
NCC	network color code. 41
NCH	Notification Channel. 41
NDC	national destination code. 40
NEF	Network Exposure Function. 183, 185, 186, 190, 192
NF	network function. 169, 171, 183–186, 190, 192, 193, 237
NFV	network function virtualization. 5, 156, 236

NG-RAN New Generation RAN. 169, 172, 176, 179–183, 185, 186, 188, 191–193
NGMN Next Generation Mobile Networks. 196
NLoS non-line-of-sight. 59
NMC network management center. 39
NMT Nordic Mobile Telephone. 1, 2, 15, 25
NOMA non-orthogonal multiple access. 160, 162, 163, 223, 236
NR New Radio. 155, 156, 169, 171–179, 182, 183, 192, 193, 236, 237
NR-U NR Unlicensed. 179
NRF Network Repository Function. 185, 186, 190, 192
NSA Non-Standalone. 156, 169–171, 192, 193, 236
NSI Network Slice Instance. 169, 171, 182, 185
NSS Network and Switching Subsystem. 30, 37–39, 233
NSSAI Network Slice Selection Assistance Information. 185
NSSF Network Slice Selection Function. 185, 186
Nt-SAP Notification SAP. 94
NWDAF Network Data Analytics Function. 185, 186

O-RAN Open RAN. 180, 181
OFDM orthogonal frequency division multiplexing. 231, 234–236
OFDM orthogonal frequency-division multiplexing. 2, 4, 56, 60, 64, 111, 114, 115, 123, 125, 127, 130, 134, 174
OFDMA orthogonal frequency-division multiple access. 2, 60, 108, 115, 116, 123, 137, 147, 162, 168, 223, 234, 235
OMA orthogonal multiple access. 162, 223
OMC Operations and Mainenance Center. 30, 39
OMSS Operation and Maintenance Sybsystem. 39
OPEX operational expenditure. 122, 123, 151, 224
OQPSK offset QPSK. 41
OSI Open Systems Interconnection. 90, 137
OSS Operation and Support Subsystem. 30, 37, 39, 233
OTDOA Observed Time Difference of Arrival. 101
OVSF orthogonal variable spreading factor. 74, 77
OWC optical wireless communications. 215

P-CCPCH Primary Common Control Physical Channel. 75
P-CSCF Proxy-CSCF. 103
P-GW Packet Data Network Gateway. 130, 235
P2P point-to-point. 183, 185
PAM pulse amplitude modulation. 77
PAPR peak-to-average power ratio. 116, 130, 174, 235
PBCH Physical Broadcast Channel. 175
PCCH Paging Control Channel. 72, 73
PCF Policy Control Function. 185, 186
PCFICH Physical Control Format Indicator Channel. 127
PCH Paging Channel. 41, 73–75, 80, 83, 95, 174
PCI precoding control indication. 87
PCRF Policy and Charging Resource Function. 130
PDC Personal Digital Cellular. 2, 3, 24, 27
PDCCH Physical Downlink Control Channel. 127, 137, 138, 175
PDCCH Physical Downlink Shared Channel. 175
PDCP Packet Data Convergence Protocol. 90, 92, 93, 129, 136, 137, 147, 175, 235, 237
PDSCH Physical Downlink Shared Channel. 134, 137

PDU	Protocol Data Unit. 91, 92, 185
PEI	Permanent Equipment Identifier. 185
PFD	Packet Flow Description. 185
PhCH	physical channel. 83, 84, 87
PHICH	Physical Hybrid ARQ Indicator Channel. 127
PHY	physical. xii, 60, 61, 83, 84, 86, 88–90, 97, 127, 130, 133, 136, 137, 143, 147, 174, 181, 227, 235
PICH	Page Indicator Channel. 75, 88
PIN	personal identification number. 45, 51, 189, 233
PLMN	Public Land Mobile Network. 37–39, 41, 46, 47, 94, 102, 140, 157, 185, 190, 192, 233
PMCH	Physical Multicast Channel. 127
PMD	Pseudonym Mediation Device. 102
PNFE	Paging and Notification Control Functional Entity. 93, 94
PPR	Privacy Profile Register. 102
PRACH	Physical Random Access Channel. 75, 77, 79, 89, 173
PRB	physical resource block. 138, 139
PS	packet switched. 92, 97, 141, 188
PSTN	Public Switched Telephone Network. 13, 14, 25, 56
PT-RS	Phase Tracking Reference Signal. 175
PUCCH	Physical Uplink Control Channel. 127, 137, 175
PUK	PIN unblocking key. 45, 51, 233
PUSCH	Physical Uplink Shared Channel. 127, 134, 137, 138, 175
QAM	quadrature amplitude modulation. 3, 34, 77, 84, 85, 134, 174
QCI	QoS class identifier. 138, 139
QoE	quality of experience. 207, 221
QoS	quality of service. 56, 57, 60, 119, 121, 125, 129, 130, 137–139, 175, 188, 199, 202, 203, 207, 212, 226, 229, 234, 237
QPSK	quadrature phase shift keying. 31, 41, 58, 68, 77, 86
RA	routing area. 141, 146
RAB	radio access bearer. 92
RACH	Random Access Channel. 41, 42, 46, 48, 73–75, 80, 83, 88, 89, 91, 94
RAN	Radio Access Network. 3, 5, 63, 65, 71, 101, 102, 106, 108, 125, 128, 129, 141, 169, 171–173, 179–185, 189, 192, 193, 237
RAND	Random Number. 45–47
RAT	radio access technology. 99, 140, 141, 145, 177, 182, 188
RB	radio bearer. 92
RCR	*Research and Development Center for Radio Systems*. 3
Rev.	Revision. 3, 4
RF	radio frequency. 38, 57, 84, 172, 174, 215–217, 221, 222, 235
RFE	Routing Function Entity. 93
RIT	Radio Interface Technologies. 153, 157
RLC	radio link control. 31, 89, 91–93, 97, 129, 136, 137, 147, 180, 182, 235
RN	relay node. 134
RNC	radio network controller. 66, 72, 93, 94, 96, 97, 106, 129, 130, 141, 146, 234
RPE-LTP	Regular Pulse Excitation - Long Term Prediction. 43, 233
RRC	root-raised cosine. 77, 78, 84, 88, 90, 91, 93–96, 127, 136, 137, 140–142, 145, 146, 175, 176, 180, 181, 188, 237
RRM	radio resource management. 38, 129, 136, 137, 139
RRU	Remote Radio Unit. 180
RS	reference signal. 175

RSRP reference signal receive power. 140
RTT radio transmission technology. 55

S-CCPCH Secondary Common Control Physical Channel. 75, 77, 87, 88
S-CSCF Serving-CSCF. 103
S-GW Serving Gateway. 130, 235
SA Standalone. 156, 169, 171, 193, 236
SACCH Slow Associated Control Channel. 41
SAP Service Access Point. 92–94
SBA Service Based Architecture. 183–185, 189, 190, 192, 193, 237
SC-FDMA single-carrier frequency-division multiple access. 108, 116, 117, 123, 130, 138, 147, 235
SCFE Shared Control Functional Entity. 93
SCG Secondary Cell Group. 177
SCH Synchronization Channel. 41, 42, 75–77, 79, 80, 87, 127, 174
SDAP Service Data Adaption Protocol. 175, 237
SDCCH Stand-alone Dedicated Control Channel. 41
SDMA space-division multiple access. 115, 223
SDN software-defined network. 5, 156, 236
SDU service data unit. 92
SEPP Security Edge Protection Proxy. 185, 186, 190, 192, 193, 237
SF spreading factor. 74
SFBC space-frequency block coding. 112
SFR soft frequency reuse. 133
SGSN Serving GPRS Support Node. 30, 72, 96–98, 102, 104, 234
SIC successive interference cancellation. 162, 223, 224
SIDF Subscription Identifier Deconcealing Function. 190, 192
SIM Subscriber Identity Module. 30, 37–40, 45, 47, 72, 96, 101, 233, 237
SINR signal-to-interference-plus-noise ratio. 65, 138
SIP Session Initiation Protocol. 103
SIR signal-to-interference ratio. 63, 88
SM-CP Short Message Control Protocol. 185
SM-RP Short Message Relay Protocol. 185
SMF Session Management Function. 172, 185, 186, 193, 237
SMS Short Message Service. 3, 25, 26, 29, 32, 37, 74, 185, 231, 234
SMSF Short Message Service Function. 185
SN subscriber number. 40
SNR serial number. 39
SNR signal-to-noise ratio. 42, 68, 85, 113, 118, 121, 134, 161
SON self-organizing networks. 110, 135, 184
SP spare. 39
SPC signaling point code. 41
SPP surface plasmon polariton. 216
SRES Signature Response. 45–47
SRIT Sets of Radio Interface Technologies. 153, 157
SRNS Serving Radio Network Subsystem. 93, 95
SRS Sounding Reference Signal. 175
SS7 Signalling System No. 7. 25, 48
STBC space-time block coding. 112
SU-MIMO single-user multiple-input multiple-output. 131, 161
SUCI Subscription Concealed Identifier. 190–192
SUPI Subscription Permanent Identifier. 189, 190, 192

TA	tracking area. 141
TA	terminal adapter. 38, 146
TAC	type approval code. 39
TACS	Total Access Communications System. 1, 2, 12, 15, 25
TAL	TA List. 141
TAU	TA Update. 141
TCH	full rate TCH. 43
TCH	traffic channel. 41–43
TD-CDMA	Time Division CDMA. 71
TD-LTE	Time Division-Long Term Evolution. 108
TD-SCDMA	Time Division Synchronous CDMA. 55, 59, 61, 69, 71, 108, 143
TDD	time-division duplex. 21, 42, 55, 56, 71, 72, 108, 109, 125, 133, 143, 145, 158, 161, 162, 174, 218, 233, 236
TDMA	time-division multiple access. 2, 3, 20, 24, 25, 27, 32, 33, 42, 43, 56, 60, 115, 162, 168, 233, 234
TE	terminal equipment. 37, 38
TFC	Transport Format Combination. 81, 82, 84
TFCI	TFC index. 84
TFS	Transport Format Set. 81, 84
TFT	traffic flow template. 138, 139
THz	terahertz. vii, 5, 199, 213–217, 230, 231, 238
TIA	*Telecommunications Industry Association*. 3, 32, 57, 60
TMSI	temporary mobile subscriber identity. 40, 45–47, 50, 233
TPC	Transmit Power Control. 87, 88
Tr	transparent mode. 91, 94
TrCH	transport channel. 73, 80, 81, 83–85, 174
TRS	Tracking Reference Signal. 175
TSDSI	*Telecommunications Standards Development Society, India*. 55, 56, 153
TTA	*Telecommunications Technology Association*. 55–57
TTC	*Telecommunication Technology Committee*. 55–57
TTD	true-time-delay. 217
TTI	transmission time interval. 58, 59, 62–64, 80–82, 84, 137, 138, 173
TV	television. 158
U-TDOA	Uplink Time Difference of Arrival. 101
UAC	Unified Access Control. 188
UAV	unmanned aerial vehicle. 197, 225
UCI	Uplink Control Information. 127, 174
UDM	Unified Data Management. 185, 186, 190, 192, 193, 237
UDR	Unified Data Repository. 185, 186
UDSF	Unstructured Data Storage Function. 185, 186
UE	user equipment. 4, 42, 57, 62, 63, 72–74, 76, 81, 84, 86–89, 91–96, 99, 102, 128–131, 134, 137–141, 143, 145, 146, 158, 172, 174–177, 185, 186, 188–191, 218, 235–237
UEA	UMTS Encryption Algorithm. 96
UIA	UMTS Integrity Algorithm. 96
UL	uplink. 3, 4, 38, 41, 42, 47, 51, 71–75, 79–84, 87, 88, 96, 97, 126, 127, 133, 137, 138, 141, 143, 145, 146, 158, 174–176, 178, 181, 233
UM	unacknowledged mode. 91, 93, 94, 97
UMB	Ultra Mobile Broadband. 59, 61, 64, 108
UMMIMO	ultra-massive multi-input multi-output. 216, 217

UMTS	Universal Mobile Telecommunications System. 1–4, 55, 57, 61, 69, 71–77, 79–82, 84, 87, 89–99, 101, 102, 104, 106, 107, 125, 127, 130, 134, 136–138, 140–143, 146, 147, 188, 231, 234, 235
UP	user plane. 90, 94, 174, 175, 177, 181, 183, 185, 189–191
UPF	User Plane Function. 172, 185, 186
URA	UTRAN Registration Area. 95
URLLC	ultra-reliable low latency communication. 5, 152, 156, 157, 167, 173, 175, 181–183, 200, 231, 236
USA	United States of America. 3, 38
USIM	Universal SIM. 72, 96–98, 190, 191, 237
UTRA	UMTS Terrestrial Radio Access. 55, 71, 72, 88, 101, 108, 125, 234
UTRAN	UMTS Terrestrial Radio Access Network. 57, 71, 72, 87–89, 93, 95, 128, 129, 137, 141, 142, 145–147, 234, 235
UWC-136	Universal Wireless Communications-136. 56, 60
V-RAN	Virtualized RAN. 180, 181
V2I	vehicle-to-infrastructure. 202
V2V	vehicle-to-vehicle. 110, 202
V2V	vehicle-to-everything. 110, 157
VAD	voice activity detection. 35
VLC	visible light communications. 215
VLR	Visitor Location Register. 30, 39–41, 45–48, 72, 96–98
VoIP	voice over Internet protocol. 58, 103, 107, 235
VPLMN	visited PLMN. 190
VR	virtual reality. 195, 197, 199–201, 203
WAP	Wireless Application Protocol. 28
WCDMA	Wideband CDMA. 3, 32, 55, 58, 61, 64, 69, 71, 74, 108, 146, 231, 234
WiMAX	Worldwide Interoperability for Microwave Access. 60, 69, 107, 110, 111, 114, 123, 231
WMAN	Wireless Metropolitan Area Network. 59, 60, 231
WP	Working Party. 196, 203, 208, 211
WRC	*World Radiocommunication Conference*. 54
XR	extended reality. 195, 198, 201, 204, 211, 213

1 Standards History of Cellular Communications Systems

1.1 Legacy Cellular Standards: From 1G to 4G

Since the first commercial deployment, standardized cellular networks have been rapidly evolving for decades. Various standards have been established for cellular technologies, which are generally classified into different generations, from First Generation (1G) to Fourth Generation (4G). This classification, however, can be usually confusing for its numerous variations in practical use. For instance, the Enhanced Data Rates for GSM Evolution (EDGE) is sometimes not only recognized as 2.5G (Garg, 2010) but also occasionally called Third Generation (3G) (Osseiran et al., 2016). Similarly, the *3rd Generation Partnership Project* (3GPP) Release 8 of Long-Term Evolution (LTE) standard is uniformly considered as 4G (Dahlman et al., 2013) but classified as 3.75G or 3.9G/3.95G in some literature (Osseiran et al., 2016; Parikh & Basu, 2011). Depending on the reference to LTE, the recognition to Long-Term Evolution Advanced (LTE-A) (Release 10) and LTE-A Pro (Release 13/14) can also vary between 4G and 4.5G, and between 4.5G and 4.9G, respectively.

The reason for such chaotic classification is rooted in the conflict between technical specifications and commercial marketing considerations. While the Second Generation (2G) systems, which are the first digital cellular systems, can be easily distinguished from the analog 1G systems, the strict performance metrics to identify 3G and 4G systems are regulated by the *International Telecommunication Union, Radiocommunication Sector* (ITU-R), in its publicly released recommendations—International Mobile Telecommunications (IMT), including IMT-2000 and IMT-Advanced, respectively. Technically, they were the Universal Mobile Telecommunications System (UMTS) and LTE-A that first fulfilled IMT-2000 and IMT-Advanced requirements, respectively. Nevertheless, since EDGE and LTE were developed as the precursors to their correspondingly next-generation cellular networks, mobile network operators (MNOs) have advertised them as 3G/4G technologies for marketing purpose, which have been thereafter widely accepted by the public and even by the wireless community.

It shall be noted that, though the commercial deployment of cellular mobile networks has been dominated by 3GPP standards, there have been other cellular standards established by other organizations suchlike *3rd Generation Partnership Project 2* (3GPP2) and *Institute of Electrical and Electronics Engineers* (IEEE), which also fulfill the performance requirements outlined by ITU-R and can be categorized into the generations of cellular technologies, as listed in Table 1.1.

It can be the most intuitive observation on the evolution of mobile networks that data rate has explosively increased from around 2.4 kbps[1] of Advanced Mobile Phone System (AMPS) to 3 Gbps of LTE-A Pro. Behind such a dramatic increase, there is a deep and fundamental change of networking technologies in every new generation, as summarized in Table 1.2.

1.1.1 1G and 2G Standards

Dating back to the 1970s, different analog 1G standards were developed and deployed in their own regions. This includes the AMPS in the Americas, the C-Netz in Germany, Portugal, and South Africa, NMT in Nordic countries, and the TACS in the United Kingdom. They mostly apply analog frequency modulation (FM) for the speech signals and frequency-shift keying

[1] This is the equivalent level, since it was analog voice signals transmitted in 1G systems.

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5_1

Table 1.1 Legacy cellular standards from 1G to 4G

1G		AMPS
		Radio Telephone Network C (C-Netz)
		Nordic Mobile Telephone (NMT)
		Total Access Communications System (TACS)
2G	2G	Global System for Mobile Communications (GSM)
		IS-95A
		IS-136
		Personal Digital Cellular (PDC)
	2.5G	General Packet Radio Service (GPRS)
		High Speed Circuit-Switched Data (HSCSD)
		IS-95B
	2.75G	EDGE
3G	3G	UMTS
		Code-Division Multiple Access (CDMA) 2000
		IEEE 802.16e
		Time-Division Synchronous CDMA (CDMA)
	3.5G/3.75G	High Speed Downlink Packet Access (HSDPA)
		Evolved High Speed Packet Access (HSPA+)
		1x Evolution-Data Optimized (EV-DO)
		EVolution-Data Voice (EV-DV)
	3.9G/3.95G	LTE Release 8
4G	4G	LTE-A
		IEEE 802.16 m
	4.5G	LTE-A Pro

Table 1.2 The data rate increase and feature upgrade along the wireless evolution from 1G to 4G

Technology	Peak data rate	Key features
1G	equiv. to 2.4 kbps	Analog system
		Frequency-division multiple access (FDMA)
		Basic voice service
2G	9.6–256 kbps	Digital transmission and power control
		Circuit switching with packet switching extension
		Time-division multiple access (TDMA)/FDMA hybrid
		Voice and basic data services
3G	384 kbps to 326 Mbps	Hierarchical cell structure
		Circuit/packet switching hybrid
		CDMA[a]
		Voice and some data services
4G	1–3 Gbps	IP-based protocols
		Pure packet switching
		Orthogonal frequency-division multiplexing (OFDM)
		Multiple-input multiple-output (MIMO)
		Mobile broadband data services

[a] Except for LTE Release 8, which uses OFDMA and is commercially accepted as 4G

(FSK) for the digital signaling (as an exception, C-Netz works with phase-modulated speech signals). They generally apply FDMA.

The 2G standards were overwhelmingly dominated by the GSM, a hybrid TDMA/FDMA digital network. Its development was initiated in 1982 by the *European Conference of Postal and Telecommunications Administrations* (CEPT), aiming at creating a common digital voice telephony network that allows international roaming across Europe. GSM was

internationally deployed since 1991. With the digital transmission and switching technologies adopted, it managed to significantly improve the voice quality, raise the network capacity, and provide supplementary low-rate data services.

Afterward, the GPRS extension, known as the 2.5G standard, was integrated into the GSM network switching subsystem to provide a packet-switched data service, with which the available network capacity is shared among many users so that the waste of bandwidth is decreased to a low level. In comparison to circuit switching, packet switching is more efficient with bandwidth utilization (Heine & Sagkob, 2003), which enhances the data capacity of Global System for Mobile Communications (GSM) networks and enables new services such as Short Message Service (SMS) and multimedia. Packet switching is considered as a turning point in the history of cellular communication, which opened the door for research and development of 3G and beyond technologies. A comparison between packet switching and circuit switching techniques along with their pros and cons is thoroughly described in Heine and Sagkob (2003); Walke (2013).

A further GSM evolution was then raised by the so-called 2.75G standard, EDGE, and its associated packet data component, Enhanced General Packet Radio Service (EGPRS). This mainly involves with addition of higher-order modulation and coding schemes. Another 2.5G option for GSM evolution was the HSCSD, which is less deployed than GPRS due to its relatively high cost. HSCSD is also an optional module in EDGE.

In parallel with GSM, there have been other regional 2G standards. For instance, in the United States of America (USA), *Telecommunications Industry Association* (TIA) and *Electronic Industries Alliance* (EIA) established the TDMA-based TIA/EIA-136 and the CDMA-based TIA/EIA IS-95A (a.k.a. cdmaOne), which were under accreditation of *American National Standards Institute* (ANSI). In Japan, the TDMA-based PDC was established by the *Research and Development Center for Radio Systems* (RCR), which became later the *Association of Radio Industries and Businesses* (ARIB). Especially, the IS-95A has its own 2.5G evolution, the IS-95B, which provides a higher data rate service through code aggregation. Usually, IS-95A and IS-95B are jointly referred to as the TIA/EIA-95, a.k.a. the CDMA One (Osseiran et al., 2016).

1.1.2 3G and 3G Evolution Standards

It was only soon after the commercial deployment of 2G when the industry began to sketch the roadmap toward 3G according to the ITU-R's IMT-2000 standards. The standardization of IMT-2000 was also a continuously evolving process that dates back to 1985 (Fukuda et al., 2002), wherein the most significant event happened in 1998, when *European Telecommunications Standards Institute* (ETSI) released the UMTS. As the first and major 3G standard matching IMT-2000 requirements, UMTS adopted CDMA in two variants, namely the Wideband CDMA (W-CDMA) and the Time-Division CDMA (TD-CDMA). A third alternative, the Time-Division Synchronous CDMA (TD-SCDMA), was later developed and deployed in China. Meanwhile in the USA, TIA also evolved IS-95 into its 3G version, the IS-2000, a.k.a. CDMA 2000. In addition, the *Institute of Electrical and Electronics Engineers* (IEEE) also established in 1999 a working group to develop standards for broadband wireless metropolitan area networks, which generated the standard family IEEE 802.16, a.k.a. the WirelessMAN. The IEEE 802.16e proposed in 2005 is the first 3G specification in this family, based on which the well-known WiMAX standards were developed for commercial implementations (Pareit et al., 2012).

To further evolve UMTS, the 3GPP was launched in December 1998, which has been one of the most important standardization organizations in cellular communications since then. Its standards inherit the legacy GSM/UMTS standards and are officially named as "Releases." Based on WCDMA, 3GPP proposed a new Radio Access Network (RAN) approach called High Speed Packet Access (HSPA). HSPA is a combination of HSDPA and High Speed Uplink Packet Access (HSUPA), which were defined in 3GPP Release 5 and Release 6, respectively. By deploying shared channel transmission, HSPA was able to enhance the packet data rate in both downlink (DL) and uplink (UL) while reducing the latency as well. In 3GPP Release 7, a further improved data rate up to 42.2 Mbps in the DL and 22 Mbps in the UL was achieved by introducing multiple-input multiple-output (MIMO) and higher-order modulation (64 quadrature amplitude modulation (QAM)), known as the HSPA+.

Another RAN approach of 3G evolution, the EV-DO, was based on CDMA 2000, within the framework of the Third-Generation Partner Project 2 (3GPP2).[2] EV-DO started with its Release 0 in 1999, which enhances the CDMA 2000 design with fast channel estimation feedback, dual receiver antenna diversity, multi-user diversity, adaptive modulation, automatic repeat request (ARQ), and turbo coding. These new technologies grant EV-DO with up to 2.46 Mbps data rate in the DL. This Release 0 was thereafter upgraded to its Revision (Rev.) A, which introduces new coding rates to further increase the

[2] Remark that it shall be distinguished from 3GPP.

peak DL rate to 3.1 Mbps. A multi-carrier evolution was added in EV-DO Rev. B, which provides higher data rates per carrier and the capability of deploying multiple carriers for the same link.

One of the most significant architectural advances of 3G technologies in comparison with 2G is the adoption of hierarchical cell structure (HCS). In 2G systems, every cell is of the same type and specified to a fixed traffic capacity, and the procedures of traffic management and handover are operated by each cell independently within its own coverage. Any attempt to increase the capacity of a certain cell will lead to an expensive cellular reconfiguration such as cell splitting and cell sectorization (Garg, 2010). To fulfill the increasing requirement of flexibility and capacity of data transmission, 3G systems generally deploy cells of multiple types with various characteristics and overlay them discontinuously in a hierarchical multi-layer topology. High-layer cells (femto/pico/micro-cells) with smaller coverage are capable of supporting higher transmission rates and denser traffic in scenarios with low mobility and small delay spread, e.g., the indoor environment. They are laid over the lower layer cells (macro-/supermacro-cells) that support higher mobility in wider coverage but with limited traffic densities. user equipment (UE) will be registered flexibly to the most appropriate cell with respect to the user profile, the service expectation, and the traffic environment.

Another identifying feature of 3G architecture is the packet-switched core network. As aforementioned, packet switching technologies had been introduced into GSM networks by the 2.5G GPRS extension. The 3G UMTS systems have inherited the GPRS topology, providing a hybrid of circuit-switched and packet-switched core networks. In comparison, the EV-DO system, focusing on data service alone as its name suggests, has applied a pure packet-switched architecture (with the voice service being provided by the legacy CDMA 2000 system alongside).

1.1.3 Precursor-to-4G Standards

Till HSPA+ (Release 7), all 3GPP releases are backward compatible with the UMTS systems. However, the LTE proposed by 3GPP Release 8 became a turning point, where OFDM is applied to replace the CDMA air interface, and a new core network called Evolved Packet Core (EPC) supersedes the UMTS backbone network. Due to these significant evolutions, LTE is no longer backward compatible with UMTS. As aforementioned, being developed as a precursor to 4G, LTE managed to raise the peak DL data rate to the level of 326 Mbps, closely approaching to the 1 Gbps requirement of 4G defined by IMT-Advanced.

Aiming at the 4G era, 3GPP2 also proposed their evolution to EV-DO, which was initially named EV-DV and later renamed into EV-DO Rev. C. Similar to LTE, EV-DV also adopts OFDM to avoid the intrinsic disadvantages of CDMA systems. Deploying frequency-division duplex (FDD), Internet Protocol (IP)-based architecture, MIMO, multi-carrier, and interference cancellation, EV-DV offers to further increase the network capacity in comparison to EV-DO Rev. B. However, it finally turned out to lose the competition against LTE, and its development was terminated at the end of 2008.

1.1.4 4G Standards

While 3GPP2 was abandoned, the LTE architecture has continuously evolved within the framework of 3GPP. By the release of LTE-A (3GPP Release 10), which added higher-order MIMO and carrier aggregation, the official 4G expectation in IMT-Advanced was for the first time fulfilled: LTE Release 10 is capable to provide up to 3 Gbps data rate in the DL and 1.5 Gbps in the UL. A set of further improvements is afterward realized by the following LTE releases (Release 11–13); this involves with enhancements in carrier aggregation, relaying, interference cancellation, higher MIMO order (Osseiran et al., 2016), and new solutions such as LTE-M (Ericsson and Networks, 2014) and narrow-band Internet of Things (NB-IoT) (Qualcomm Incorporated, 2015) to support machine-type communication (MTC).

Meanwhile, in 2011 IEEE also published its 4G-approved WiMAX standard, the IEEE 802.16m, a.k.a. WirelessMAN-Advanced (Pareit et al., 2012). However, it is significantly less accepted in comparison to the LTE-A.

1.2 The 5G Era

Driven by the exploding demand for faster, more reliable, and ubiquitous wireless services, further evolution of cellular networks toward Fifth Generation (5G) and beyond has been undergoing, mainly in the framework of 3GPP.[3] Similar to the

[3] Besides that, IEEE also proposes its own plan of 5G development.

3G and 4G standards, the 5G requirements of technical performance are defined by ITU-R in its IMT-2020 standard, which calls for dramatic enhancement to cellular services in various aspects. Furthermore, empowered by the emerging technologies of network function virtualization (NFV), software-defined network (SDN), and network slicing, 5G is expected to deliver heterogeneous services on top of a shared network infrastructure. IMT-2020 therefore defines three different reference use scenarios, namely enhanced mobile broadband (eMBB), massive machine-type communication (mMTC), and ultra-reliable low latency communication (URLLC), respectively.

The world's first full set of 5G standards is the 3GPP Release 15, often informally called "5G Phase 1," which was initiated in 2016 and closed in 2019. It mainly focuses on eMBB services and basic blocks that build the new 5G system, describing the 5G RAN, a.k.a. New Generation RAN (RAN), and the new 5G Core (5GC). Following that, Release 16, a.k.a. "5G Phase 2", which started at the end of 2017 and ended in July 2020, focuses on the use scenarios of mMTC and URLLC.[4] A more detailed overview of the standardization efforts of 5G will be provided in Chap. 9, and some highlights of Release 15 will be presented in Chap. 10.

1.3 The Road Towards 6G

As of today, 5G has been rolled out in many countries, and the number of 5G subscribers has quickly grown to 521 million globally by the end of 2021. Meanwhile, societal needs have still been increasing and will continue to increase prolifically in the next decade, which will give rise to diverse use cases that cannot be served under the umbrella of 5G networks. Furthermore, the upcoming evolution of Industry 4.0 to its next generation will also shepherd the way for a surge in connectivity density beyond the limits of what 5G promises (Wu et al., 2021a). Recognizing these emerging needs, both academia and industry have begun to shift their focus toward the next generation, known as Sixth Generation (6G) or IMT-2030. The ITU-R is actively working on the IMT-2030 standard, set to shape the future of 6G. This includes studies and evaluation of communications over 100 GHz, a crucial frequency band for 6G technology. The efforts of ITU-R, alongside those of the 3GPP with its "6G Basic" standards starting from Release 21, signify a global and collaborative approach toward developing the technical framework for 6G.

As the technological landscape evolves toward 6G, several key enabling technologies emerge as critical pillars. These include the expansion into THz frequency bands, offering significantly higher data rates and capacity, and the advancement in AI and machine learning algorithms for intelligent network management and optimization. Furthermore, the integration of non-terrestrial networks, including satellite and airborne communications, is essential for achieving truly global coverage and connectivity. The development of energy-efficient and high-frequency compatible materials and devices is also a cornerstone, addressing the challenges associated with higher frequency bands. These technologies not only signify a leap in communication capabilities but also lay the foundation for transformative applications, ranging from ultra-reliable, low-latency communication to enhanced mobile broadband and massive machine-type communications.

Beyond an incremental improvement of conventional key performance indicators (KPIs) such as throughput and latency, the next evolution of mobile networks is also expected to bring fundamental and revolutionary changes from the perspectives of technology principles, network architecture, and service paradigms, just as the previous evolutions did. Generally, it is envisaged that 6G will be ubiquitously covering every corner of the planet, delivering not only data but also trustworthy intelligent services with extreme performance (Yang et al., 2019; Tariq et al., 2020; Liu et al., 2022). Furthermore, growing concerns about the energy crisis and climate change are compelling people to prioritize sustainability as a key 6G value (Hu et al., 2020; Feng et al., 2021; Han et al., 2021).

These opinions are fully reflected in the vision of the European Union's 6G flagship project, Hexa-X, which envisages 6G to enable a plethora of emerging use cases, connecting the physical, digital, and human worlds more seamlessly than ever before (Uusitalo et al., 2021). This convergence will not only enhance the interaction between the physical and digital realms but will also integrate human cognition and interaction into the digital sphere, creating a more immersive and interactive cyber-physical system. We will further elaborate on the ongoing research and development efforts toward the first 6G standard in Chap. 11 and examine the most widely acknowledged key enabling technologies for future 6G systems in Chap. 12.

[4] At the time of publishing this book (July 2024), 3GPP has frozen its Release 17 and is closing the Release 18 package.

1.4 Summary

This chapter delves into the intricate evolution of cellular communication systems, tracing their journey from the earliest analog networks to the cutting-edge technologies of today. Beginning with the legacy cellular standards, we explore the progression from 1G to 4G, highlighting the challenges of classification due to the interplay between technical specifications and commercial marketing. The chapter underscores the pivotal role of organizations like 3GPP, 3GPP2, and IEEE in shaping the cellular landscape and the transformative impact of packet switching technologies introduced in 2G systems. As we transition to 3G and its subsequent evolutions, the narrative emphasizes the architectural advances, such as hierarchical cell structures and packet-switched core networks, that paved the way for 4G.

The dawn of the 5G era marks a significant leap from its predecessors. Characterized by the revolutionary integration of massive MIMO technology and millimeter wave (mmWave) frequencies, 5G achieves unprecedented increases in data throughput and network capacity. This transition not only signifies a dramatic enhancement in speed and efficiency but also lays the foundation for innovative applications in the Internet of Things (IoT), smart cities, and autonomous systems, setting the stage for a deeply interconnected and intelligent global network.

With 5G deployments gaining momentum worldwide, the chapter concludes by casting a visionary gaze toward the horizon of 6G, anticipating a future where cellular networks will seamlessly intertwine the physical, digital, and human realms, driving the next wave of cyber-physical systems. Through this comprehensive overview, readers gain a deep understanding of the milestones that have marked the cellular communications journey and the innovations that promise to shape its future. From the next chapter, we will elaborate on the evolution of each generation chapter by chapter.

1.5 Exercises

1. List the key organizations involved in cellular technology standardization and briefly describe their roles.
2. In which cellular generation was packet switching introduced? Briefly explain its impact.
3. What were the key technological advancements that defined 2G cellular systems?
4. What were the key technological shifts that occurred in the transition from 3G to 4G?
5. Identify one key technological feature for each cellular generations from 1G to 5G.
6. What are some of the anticipated features of 6G networks?

Evolution to First-Generation (1G) Mobile Cellular Communications

2

2.1 Pre-Cellular Systems

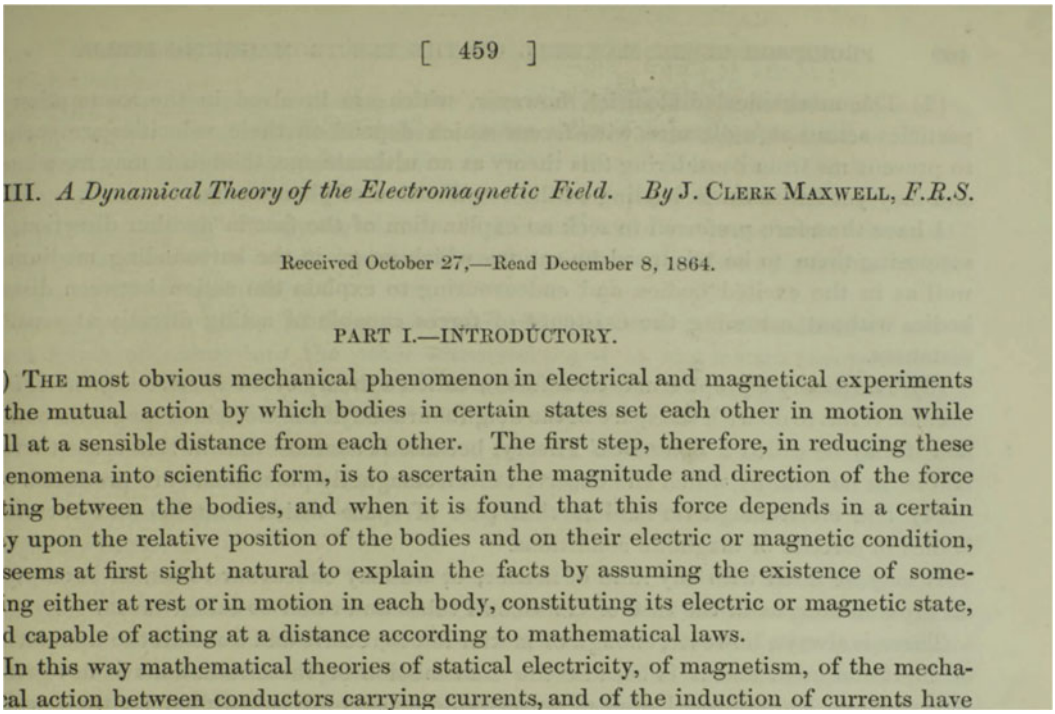
When we are talking about "wireless communications," our mind would directly turn to images of modern electrical and electronic devices used in contemporary society. Nevertheless, as John Kingman has noted, "Wireless communication is as old as Biblical times." A wide sense of "wireless communications" had been practically used for thousands of years ago, as ancient people sought to convey critical information, such as the invasion of enemies, through the use of smoke, torches, flashing mirrors, signal flares, semaphore flags, and other means. Long-range transmission was also accomplished by forwarding messages continuously along a network of relaying stations. For instance, the ancient Chinese constructed a communication system consisting of beacon towers, forts, walls, and ditches along their northern borders, as an important integral part of the Great Wall defensive systems to thwart potential invasions of nomadic groups. The beacon towers, known as *fenghuotai* in Chinese, relayed military alerts across the frontier by utilizing smoke signals during the day and torches at night. The scale of invasion was indicated by the number of smoke signals or torches. The ancient Greeks invented a method called the heliograph to convey military intelligence, which was first recorded during the Battle of Marathon in 490 BC when the Persians invaded Greece. Using mirrors to reflect the flashes of sunlight, it is considered the first wireless telegraph system.

With the development of electric technology, these infant communications systems were replaced by the wired telegraph that transfers signals over landlines. In 1837, Samuel Morse patented an electric telegraph, which conveys text messages called telegraphs through the famous Morse code consisting of dots and dashes. On March 7, 1876, Alexander Graham Bell successfully filed a patent for the telephone, and days later, he made the first-ever telephone call. The wired phone revolutionized communication by carrying information-rich voice signals. In his publication of "A Dynamical Theory of the Electromagnetic Field," as shown in Fig. 2.1, James Clerk Maxwell (1865) presented that electric and magnetic fields travel through space as waves moving at the speed of light. The unification of light and electrical phenomena led to his prediction of the existence of electromagnetic waves in 1873. In 1887, Heinrich Hertz conclusively proved the existence of electromagnetic waves by discovering radio waves.

In the summer of 1895, a few decades after the invention of the wired telegraph and telephone, Guglielmo Marconi successfully demonstrated the feasibility of wireless communications through his first experiment of radio transmission up to 2 miles. This milestone indicates the birth of wireless communications. Since then, a wide variety of communications services such as wireless telegraph, wireless telephony, radio/television broadcasting, satellite communications, and wireless local area networks, as well as non-communications applications, e.g., radar, remote sensing, and radio astronomy, were adopted across the world and substantially reshaped modern society. As the most successful category of wireless communications, mobile cellular networks experienced explosive growth in the last decades. With its ease of deployment, economic efficiency, portability, flexibility, and scalability compared to wired networks, it became one of the critical infrastructures to empower modern society and drastically reshaped human behaviors in business, education, entertainment, and personal life.

During World War II, the development of radio technology stepped into a fast track for the major countries, who wanted to gain military advantages over their enemies. In addition to the advent of radio detection and ranging (RADAR) technology, another radio communication product known as walkie-talkie played an important role on the battlefield. Motorola developed a symbolic portable phone SCR-536 for the US army, as shown in Fig. 2.2. This push-to-talk radio transceiver operates in a half-duplex mode, allowing one unit to transmit its signal while other units in its proximity listen.

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5_2



[459]

III. *A Dynamical Theory of the Electromagnetic Field.* *By* J. Clerk Maxwell, *F.R.S.*

Received October 27,—Read December 8, 1864.

PART I.—INTRODUCTORY.

) The most obvious mechanical phenomenon in electrical and magnetical experiments the mutual action by which bodies in certain states set each other in motion while ll at a sensible distance from each other. The first step, therefore, in reducing these enomena into scientific form, is to ascertain the magnitude and direction of the force ting between the bodies, and when it is found that this force depends in a certain y upon the relative position of the bodies and on their electric or magnetic condition, seems at first sight natural to explain the facts by assuming the existence of some- ng either at rest or in motion in each body, constituting its electric or magnetic state, d capable of acting at a distance according to mathematical laws.

In this way mathematical theories of statical electricity, of magnetism, of the mecha- al action between conductors carrying currents, and of the induction of currents have

Fig. 2.1 The famous paper by James Clerk Maxwell

Fig. 2.2 The symbolic portable phone Motorola SCR-536 during World War II

The whole unit weighed 2.3 kg with a range of approximately 5 km by using the AM mode operating in the frequency range from 3.5 MHz to 6.0 MHz. It was primitive, but electrical engineers at that time got much experience in developing more advanced wireless telephones.

The early mobile telephones were used before the advent of cellular technology, referred to as pre-cellular or sometimes zero generation (0G) systems. Differing from earlier closed radiotelephone systems, pre-cellular technologies, such as Mobile Telephone Service (MTS) and Improved Mobile Telephone Service (IMTS), were available as commercial service that was connected to the public switched telephone network. In 1946, Motorola in conjunction with the Bell System launched the first MTS system in the USA as an extension of the wired telephone service. On June 17, 1946, the Bell System demonstrated the world's first mobile call in the City of St. Louis. The radio transceiver weighted 36 kg was mounted in a car and thus called a car phone. In the beginning, the MTS system only provided three channels for a few subscribers in the metropolitan area. The number of channels was increased to 32 soon later over three frequency bands, i.e., 35 MHz to 44 MHz (9 channels), 152 MHz to 158 MHz (11 channels), and 454 MHz to 460 MHz (12 channels). This revolutionary service achieved great success and expanded to over 100 cities across the United States within 3 years, attracting a total of approximately 5000 users. With two major technical advances, i.e., direct dialing allowing a phone call without manual connection through a live operator, and full duplex transmission, by which two communicating parties can talk simultaneously, the successor of MTS named IMTS was rolled out in 1964.

Outside the USA, there are similar pre-cellular systems developed in major advanced countries. In 1952, West Germany opened the country's first-generation mobile telephone network referred to as A-Netz, which was upgraded to B-Netz, as the second-generation technology. As the first automatic system in Europe, the Soviet Union "Alta" mobile telephone system was fully operational in 1965. It operated in 150 MHz band initially and migrated to 330 MHz band as the network capacity grew. In 1965, Japan deployed its own mobile telephone system named Advanced Mobile Telephone System (AMTS), which was operated on the 900 MHz frequency band and used in the 1960s and 1970s before the launch of the 1G cellular system. Offentlig Landmobil Telefoni (OLT), or Public Land Mobile Telephony, was Norway's first mobile telephone system, using FM on 160 MHz to 162 MHz for the uplink and 168 MHz to 170 MHz for the downlink. The OLT network served maximally 30,000 subscribers since its establishment in 1966 and becomes the largest network in the world at that time. In 1971, Autoradiopuhelin, which means "Auto Radio Phone (ARP)," was launched as Finland's first commercial mobile telephone network. Sweden and Denmark introduced a manual mobile system operated in the 450 MHz frequency band in 1971, called Mobile Telephony System D (MTD), which got a total of 20,000 subscribers.

Early Milestones of Wireless Communication: JPL (1995)

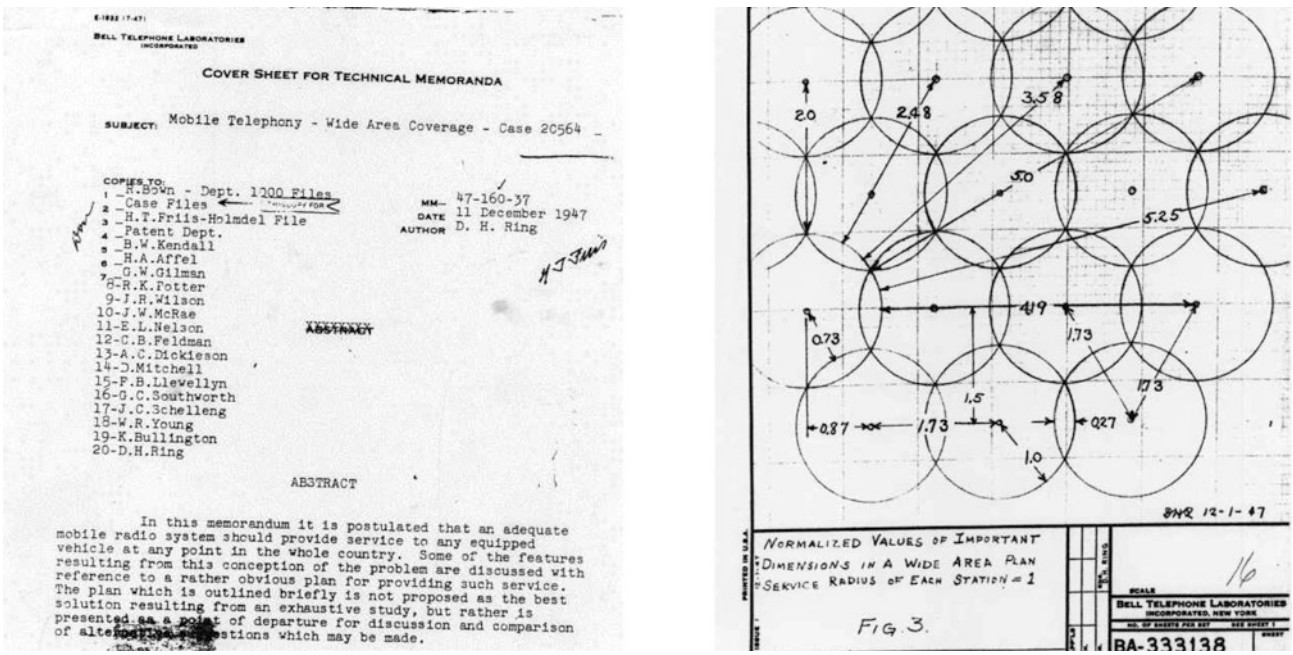
- 1864: James Clerk Maxwell predicted the existence of electromagnetic waves.
- 1887: Heinrich Hertz discovered radio waves, proving Maxwell's prediction.
- 1895: Guglielmo Marconi successfully demonstrated radio transmission.
- 1901: Marconi received a Morse message across the Atlantic.
- 1904: J.A. Fleming patented the diode.
- 1906: Lee DeForest patented the triode amplifier.
- 1907: Trans-Atlantic commercial wireless communications service.
- 1915: First wireless voice transmission.
- 1920: Marconi discovers short-wave radio.
- 1920: First commercial radio broadcasting (in Pittsburgh)
- 1921: Police car dispatch radios (Detroit).
- 1930: British Broadcasting Corporation (BBC) began television experiments.
- 1935: First telephone call around the world.
- 1946: Motorola opened the first MTS service in the USA.
- 1947: The inventions of the transistor and the cellular network layout.
- 1952: West Germany launched A-Netz.
- 1965: the Soviet Union's "Alta" system was fully operational.
- 1965: Japan deployed the AMTS mobile telephone system.
- 1966: Norway opened the OLT mobile network.
- 1971: Finland deployed the ARP mobile network.
- 1971: Sweden and Denmark introduced MTD telephone service.

2.2 The Advent of Cellular Networks

Designed for only a few mobile telephone users, these pre-cellular systems installed a central base station to cover an entire metropolitan area. The transmit power is very high, e.g., 100 Watts (W) for an IMTS base station, which is sufficient for a range of 60 km to 100 km, in comparison with less than 1 W on cutting-edge 5G small base stations. Each voice channel is exclusively dedicated to one conversation between the caller and the callee. There is no frequency reuse within a large cite, such that the licensed spectrum can accommodate only a few channels, leading to very limited system capacity. In the 1970s, before the deployment of cellular networks, a customer wishing to subscribe to mobile telephone service had to wait for up to 3 years until an incumbent subscriber terminated their mobile subscription.

Cellular Network *Small capacity of per-cellular systems cannot satisfy the rising demand for mobile telephone services, driving the advent of an elegant network design known as the cellular network. The world-renowned research institution, Bell Laboratories, accomplished two historic innovations for modern society in the same year 1947—the transistor and the cellular network.*

In 1947, William R. Young, an engineer who worked at AT&T Bell Laboratories, presented the cellular concept of the hexagonal geometry throughout a wide coverage area, as illustrated in Fig. 2.3. Douglas H. Ring, also at Bell Labs, expanded on Young's initial concept. He sketched out the basic design for a standard cellular network and published the intellectual groundwork as a technical memorandum entitled *Mobile Telephony—Wide-Area Coverage* in Bell Labs' internal journal on 11 December 1947 (Ring, 1947). Per-cellular systems used a single base station to cover a city, where the whole frequency band only supported small capacity. If the same frequency can be reused in different sites which are separated from one another by sufficient distances, a small block of spectrum can accommodate a large-scale telephone service with attention to cost restraint. In a cellular network, a wide coverage is divided into a continuum of hexagonal areas called cells. A radio station is placed in the center of a cell and is assigned a portion of the whole frequency band. Within a cluster of cells, any two adjacent cells are non-overlapped in frequency, such that precious spectral resources are reused at spatially separated sites taking advantage of the fact that the power of a transmitted signal drops dramatically with the increase of propagation distance.


BELL TELEPHONE LABORATORIES INCORPORATED

COVER SHEET FOR TECHNICAL MEMORANDA

SUBJECT: Mobile Telephony - Wide Area Coverage - Case 20564

COPIES TO:
1 - H.Brown - Dept. 1200 Files
2 - Case Files
3 - H.T.Friis-Holmdel File
4 - Patent Dept.
5 - B.W.Kendall
6 - H.A.Affel
7 - G.W.Gilman
8 - R.K.Potter
9 - J.R.Wilson
10 - J.W.McRae
11 - E.L.Nelson
12 - C.B.Feldman
13 - A.C.Dickieson
14 - D.Mitchell
15 - P.B.Llewellyn
16 - G.C.Southworth
17 - J.C.Schelleng
18 - W.R.Young
19 - K.Bullington
20 - D.H.Ring

MM 47-160-37
DATE 11 December 1947
AUTHOR D. H. Ring

ABSTRACT

ABSTRACT

In this memorandum it is postulated that an adequate mobile radio system should provide service to any equipped vehicle at any point in the whole country. Some of the features resulting from this conception of the problem are discussed with reference to a rather obvious plan for providing such service. The plan which is outlined briefly is not proposed as the best solution resulting from an exhaustive study, but rather is presented as a point of departure for discussion and comparison of alternative suggestions which may be made.

Fig. 2.3 The cover sheet and conceptual drawing of the cellular concept published by D. H. Ring on December 11, 1947 (Ring, 1947)

Despite the elegant design of the cellular concept, it was really a long journey during the development process of the cellular system from a great concept to a real network due to technological and hardware barriers. As early as 1947, the US Federal Communications Commission (FCC) issued AT&T a spectrum license to deploy the cellular services. The initial system design had been mostly completed in the 1960s. The first trial network consisting of ten cells was eventually installed in 1977, when many of the original designs were outdated (Goldsmith, 2005). Based on this trial network, Bell Labs developed the first standards of a cellular network called Advanced Mobile Phone System (AMPS) (Young, 1979). AMPS was successfully deployed in the United States, and many other countries and smoothly evolved into a second-generation digital cellular standard known as IS-54 (where IS stands for Interim Standard). In the early 1980s, with tens of years of technical development, 1G networks were finally rolled out to offer commercial mobile telephony service to the public. With its ease of deployment, economic efficiency, portability, flexibility, and scalability compared to wireline networks, mobile cellular networks experienced explosive growth in the last decades. It became one of the critical infrastructures to empower modern society and drastically reshaped human behaviors in business, education, entertainment, and personal life.

Although the United States is the birthplace of the cellular concept, the world's first commercial cellular network was not appeared there. The milestone of 1G cellular network was happened in December 1979, when the Japanese network operator Nippon Telegraph and Telephone (NTT) opened the first cellular network based on Mobile Cellular System (MCS). The initial network comprised 88 cells covering all metropolitan area districts in Tokyo (unlike IMTS where the network is just an extension of the public switched telephone network). Instead of accessing only to one radio tower, inter-cell handover of the call among different cells was realized. The network operated in frequency-division duplex (FDD) mode, where a pair of frequency bands on 900 kHz, namely 870 MHz to 885 MHz (uplink) and 925 MHz to 940 MHz (downlink), were assigned. The speech signal of each mobile subscriber was carried by an analog channel with a bandwidth of 25 kHz. Hence, the licensed spectrum of 2×25 MHz can accommodate a total of 600 pairs of voice channels, far greater than the capacity of pre-cellular systems. Within 5 years, the network was expanded to cover the entire population of Japan, making it the first country to provide a nationwide cellular communications service in the 1980s.

The NTT network is the world's first cellular network, but mobile terminals at the initial stage were still car phones. The terminal was too heavy (the original equipment commercialized by Motorola in the 1940s weighs around 36 kg), with a high demand for power supply. Hence, the radio transceiver had to be amounted into automobiles. Until 1985, several years after the launch of the cellular network, NTT released shoulder phones that were still bulky but at least can be carried freely by a person. Car phones had been in use in US cities since the 1930s, but the talent engineer Martin Cooper believed that the mobile phone should be a personal telephone—something that would represent an individual, so you could assign a number, not to a place, not to a desk, not to a home, but to a person. He led a Motorola team to develop the first cell phone prototype and successfully demonstrated the world's first cell phone call at the New York City Hilton in midtown Manhattan on April 03, 1973. After a 10-year journey of bringing the cell phone to market, Motorola published its historic product—DynaTAC (DYNamic Adaptive Total Area Coverage) 8000X in 1983—the first commercial cell phone that was lightweight (weighed 1.1 kg) and small enough (25 cm) to be hold be hand. At that time, the cell phone was quite expensive, for example, the Motorola DynaTAC 8000X was sold with the price of $3,995 in 1984, and in addition, the subscription fee for mobile call service was also very high. Hence, owning a cell phone in the 1980s was a symbol of affluence and social status. Without any doubt, Motorola was the most successful company with an exceptionally influential role in the early stage of developing cell phones and cellular communications technologies. Following its iconic DynaTAC 8000 series, the company released the world's first flip phone Motorola MicroTAC and then the first clamshell phone Motorola StarTAC, which was not only the world's smallest at the time but also the most lightweight with an extreme weight of 105 g. These early days also witnessed the rise of Nokia in mobile communications, which was successfully transformed from a company of forestry, cable, and rubber to the world's second largest cell phone maker with the launch of their Cityman series followed by the Nokia 101 candy bar design as opposed to the previous "bricks" (Linge & Sutton, 2014).

When Motorola was endeavoring to commercialize cell phones, Bell Labs developed the first cellular network standard, i.e., AMPS in the United States (Frenkiel & Schwartz, 2010). In October 1983, after several years of the launch of the NTT network in Japan, Ameritech opened the US first commercial cellular network in Chicago. Despite the later launch than other regions, a breakthrough brought by the United States was that cellular service was offered through cell phones rather than car phones. Scandinavia is the leading area in world's mobile communication industry, pioneered to develop the first European cellular standard named the Nordic Mobile Telephone (NMT). In 1981, Norway and Sweden deployed the first NMT network, followed by Denmark and Finland in the subsequent year. Due to the joint situation of multiple countries, it naturally becomes the first cellular network that can support international roaming. NMT got a big success, and the number of subscribers reached 110,000 in Scandinavia and Finland in 1985, made it the world's largest mobile network then. The NMT system was initially operated in 450 MHz (hence also known as NMT-450) and adopted a channel bandwidth of

25 kHz. As the network grew, extra frequency bands, i.e., 890 MHz to 915 MHz for the uplink and 935 MHz to 960 MHz for the downlink, were assigned in 1986. The system operated in these high-frequency bands became known as NMT-900. Until 2020, a small-scale NMT network was still in operation in Russia to offer universal communication services in some remote, sparsely populated areas taking advantage of the long-range propagation distance and good penetration capability of radio waves at the 450 MHz band. Because of the low entry threshold of technologies, 1G system has a diverse standard developed by different countries, including Total Access Communications System (TACS) first implemented by the United Kingdom in 1983, C-450 in Germany (1985), and Radiocom 2000 in France (1986). Although European countries developed many standards, the 1G European standards were incompatible due to the selection of different frequency bands, air interfaces, and networking protocols (Vriendt et al., 2002).

2.3 1G Analog Cellular Standards

To satisfy the increasing demand for mobile telephony services, leading countries started the development of cellular mobile communications systems in the 1970s or early. In addition to frequency reuse, other cutting-edge technologies at that time, including freuqnecy-division multiple access (FDMA), frequency-division duplex (FDD), and FM, were employed to provide a network capacity far larger than that of per-cellular systems. As a result, there are many different 1G standards from these countries or regions, including AMPS, NMT, TACS, MCS, C450, and Radiocom2000. These 1G standards have many similarities but still some peculiarities. This section will provide a comprehensive review of 1G analog standards, among which AMPS is regarded as a good representative that achieved great technological and commercial success. Consequently, more details about AMPS are provided to give the readers more insights into the evolution of 1G.

2.3.1 Advanced Mobile Phone System/AMPS

Mobile telephony services have been commercially available in the United States for decades since the launch of MTS in the 1940s, but customers who want to subscribe to telephone service in their cars had to wait for as long as 2 or 3 years in some cities. Until late the 1970s, the system capacity was still quite limited because each conversation occupied a voice channel at a time within a metropolitan area. The frequency of a voice channel can only be reused in other cities with a sufficient separation of around 100 km, leading to discontinuous service coverage where the areas in-between two cities have no service. Even if two service areas are adjacent, a mobile subscriber traveling from one area into another has to interrupt the conversation and initiate another call due to no capability of handover.

To meet the market demand, the *Federal Communications Commission* (FCC) allocated a spectrum of 40 MHz in the 800 MHz frequency band and invited industry proposals on a novel mobile telephone system to efficiently use the spectrum. In 1977, the FCC issued a license to Illinois Bell Telephone Company to build an experimental network, using AMPS developed by Bell Labs. Originating from the initial concept of cellular networking proposed in 1947, it underwent quite a long journey to become a practical network. The system design had been almost completed in the 1960s, followed by an extensive trial (technical and commercial) to optimize the system parameters and verify the basic planning rules for a cellular layout. In 1978, Bell Labs in cooperation with Illinois Bell Telephone Co., the American Telephone and Telegraph (AT&T) Co., and Western Electric Co constructed a large-scale and fully operational trial AMPS network. Ten cells covered approximately 2100 square miles in the city of Chicago, as well as the neighboring suburban and rural areas. The first phase of testing was an equipment test with around 100 mobile terminals, and the second phase, a service test, offered a capacity for more than 2000 users (Ehrlich, 1979).

The AMPS system aimed to substantially improve mobile telephone service compared to the pre-cellular systems, with the following major objectives:

- A remarkable increase in system capacity
- Improved voice quality comparable to that of wired phones
- A cellular structure supporting the handover of voice calls
- Efficient use of the frequency spectrum
- System scalability for the continuous growth of subscribers and traffic density
- Lower service cost

In 1983, the FCC finally issued commercial operation licenses for analog cellular networks. In addition to an initial spectrum of 40 MHz, an additional 10 MHz spectrum was allocated to enable the capacity expansion to 832 channels over 50 MHz spectrum. AMPS employed FDD to separate downlink and uplink signals, where the transmission from the mobile stations to base stations used the frequency band from 824 MHz to 849 MHz, while the 869 MHz to 894 MHz band was applied for the transmission from the base stations to mobile stations. Spectral sharing particularly termed multiple access in wireless communications is implemented by multiplexing the signaling dimensions along the time, frequency, code, or space domain. AMPS adopted FDMA to divide the whole frequency band into a parallel of non-overlapping channels, each of which has a bandwidth of 30 kHz. With this spacing, 832 pairs of channels can be allocated out of a 50 MHz frequency band. For a highly competitive market, the FCC issued two licensees known as A and B carriers with different frequency channels within each geographical area. Each carrier within a market area was assigned to a total of 416 paired channels consisting of 21 control channels and 395 voice channels. To minimize self-interference of equipment, each pair of voice channels was separated by a large margin of 45 MHz.

AMPS uses a simple analog technique called frequency modulation to modulate speech signals, with a maximal frequency deviation of 12 kHz. The voice signals are first converted into electrical signals, which are then used to modulate the frequency of the carrier wave. In 1935, Edwin Armstrong demonstrated FM for the first time, which is superior to amplitude modulation (AM) that suffers from ignition noise. FM has been the prime modulation technique used for mobile communications since the late 1930s (Goldsmith, 2005). The capture effect of FM can improve spectrum efficiency by effectively suppressing co-channel interference. If two cell sites simultaneously transmit signals, an FM receiver detecting two separate transmission peaks on the same channel will lock onto the stronger one while suppressing the interfering signal. Each control channel can be associated with a group of voice channels. Thus each set of voice channels can be split into groups of 16 channels, controlled by a different control channel. Although AMPS is an analog cellular system, control channels then were already digitized. Signaling was transmitted between a base station and mobile stations at a data rate of 10 kbps. The signaling data was digitally modulated using frequency-shift keying (FSK) and the Manchester encoding for error correction.

The AMPS infrastructure is composed of two kinds of network equipment: Base Transceiver Station (BTS) and Mobile Telephone Switching Office (MTSO), as illustrated in Fig. 2.4. A BTS is the interface between the Public Switched Telephone

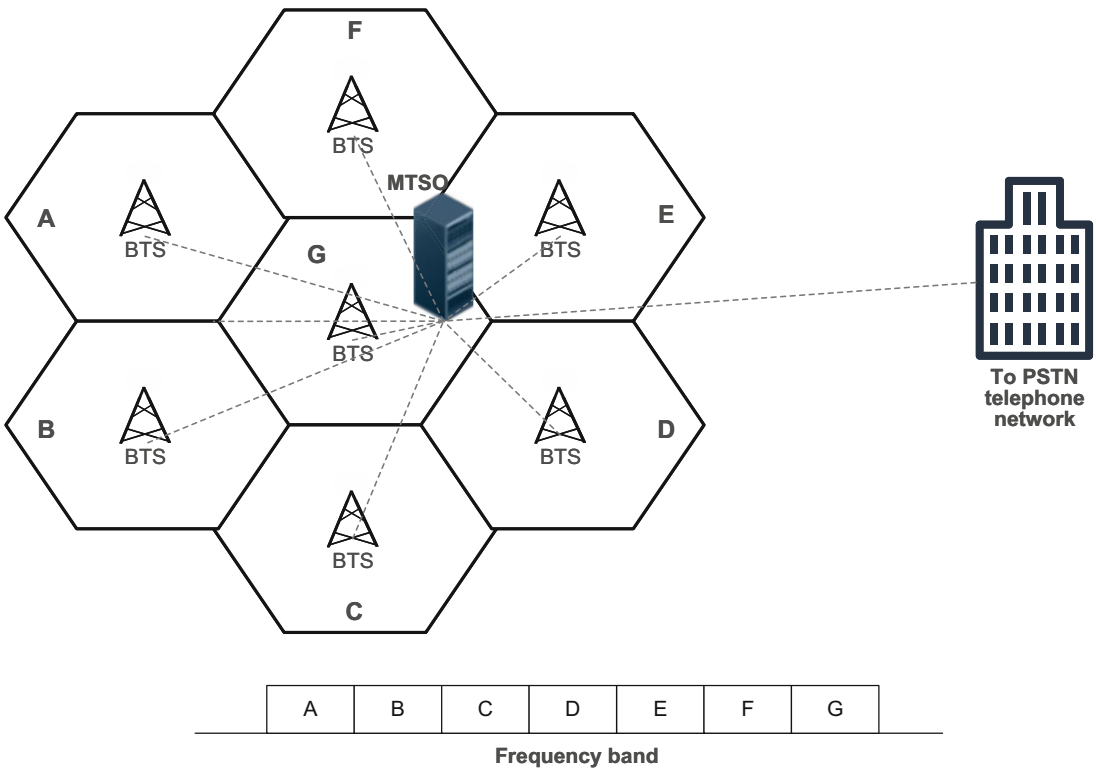


Fig. 2.4 The architecture of the AMPS system

Network (PSTN) and the wireless links to communicate with various mobile stations. For this purpose, a BTS is equipped with radio transceivers to transmit the signals toward mobile users and receive the signals from mobile users. It also has microwave or wireline transmission equipment connected to the MTSO. Small programmable controllers exist at each BTS for local processing such as call setup, call monitoring, call terminating, mobile station locating, and handover. At each cell site in the center of a cell, a BTS is mounted on the top of a building or a hill with an omnidirectional antenna to cover the circular area or 120° directional antennas for a sector. To reuse the spectrum in different sites but keep co-channel interference acceptable, the frequency band is divided into non-overlapping parts, which are assigned to a set of adjacent cell sites (called a cluster). Based on the evaluation of the performance, cost, and reliability, AMPS designers select 7 cells per cluster for directional sites and 12 cells per cluster for omnidirectional sites (Donald, 1979). The overall operation of the system is under the control of a central processor in the MTSO. The MTSO is connected via wireline or microwave links to all BTSs within a metropolitan service area by a set of voice trunks—one trunk for each radio channel—and data links, over which the MTSO and BTSs exchange messages necessary for processing calls. In addition to performing the switching function to connect the PSTN, MTSO also takes responsibility for handling the overall control of the network, allocating channels within each cell, coordinating inter-cell handover when a mobile station traverses a cell boundary, routing calls to and from mobile users, and detecting system faults.

The commercial use of AMPS not only substantially alleviated the capacity limit of mobile telephone service and consolidated a fundamental basis for the application of cellular networks, but it also suffered from many shortcomings (Frenkiel & Schwartz, 2010). First, the use of the spectrum is also not sufficiently efficient. AMPS employed a frequency reuse factor of 7 with 120° directional antennas or 12 with omnidirectional antennas to suppress strong co-channel interference. As a consequence, each cell site uses only a small portion of the total channels. The capacity of a cell was restricted to accommodate more users. Second, the designers of AMPS did not sufficiently consider the security issues, where the identification information of a mobile station, i.e., Electronic Serial Number (ESN) and Cellular Telephone Number (CTN), can be easily acquired by a fake BTS and then illegally reuse in other sites. Third, the cell phone was quite expensive at that time and the subscription fee for mobile call service was also very high, blocking its massive use in the public. Despite these limitations, AMPS played a significant role in the development of cellular communication systems and served as the basis for several other cellular standards around the world, such as the TACS system in the UK, the NMT system in Nordic countries, and the JTACS system in Japan.

The continuous rise of market demand and increasingly high expectations of better mobile telephone service imposed the necessity for a more capable, efficient, affordable, secure, and robust cellular technology. The successor of AMPS was a digital cellular system called Digital Advanced Mobile Phone System (D-AMPS), commercially deployed since 1993 in the United States and some other countries. With the support of digital components, a more advanced technology called time-division multiple access (TDMA) was employed. The digitized speech signal is first compressed such that each 30 kHz channel can simultaneously accommodate three digital voice users. To facilitate a smooth upgrade from 1G analog to 2G digital systems, D-AMPS inherited the system architecture and signaling protocol from its predecessor. A new mobile station integrated with a D-AMPS module can initially access the network via a legacy AMPS control channel and then request a digital voice channel in the areas where the D-AMPS service is already opened. Otherwise, it can still operate in the analog mode where the D-AMPS service is not available. D-AMPS was also named IS-54 and later IS-136 by the Electronics Industries Association and Telecommunication Industries Association (EIA/TIA), as the first American 2G digital cellular standard.

2.3.2 Nordic Mobile Telephone/NMT

Nordic countries are important players in the development of mobile communications, and it may be no coincidence that two Nordic companies, Nokia and Ericsson, dominated the world mobile market for a long time. Three pre-cellular mobile telephone systems have been deployed there: ARP – Auto Radio Phone in Finland, MTD – Mobile Telephony System in Sweden and Denmark, and OLT - Public Land Mobile Telephony in Norway. To meet the rising market demand for mobile voice traffic, which already saturated the pre-cellular systems, Nordic Telecommunications Administrations specified the NMT standard as a compatible cellular system in Nordic countries. The key technologies were ready by 1973, and the specifications for base stations were completed in 1977. Although NMT was a Nordic invention, Ericsson built the first commercial network in Saudi Arabia in September 1981, as a pilot test project. A few months later, the NMT service was offered in Norway and Sweden, followed by Denmark, Finland, and Iceland in the subsequent year. By 1985, the number of subscribers in the Nordic region reached 110,000 in Scandinavia and Finland, making it the world's largest mobile network at that time. Many European and Asian countries then adopted the same system.

NMT standard has two variants based on the operating frequency bands: NMT-450 and NMT-900. Using the FDD mode, the signal transmission from the motile terminals to the cell sites was assigned to the frequency band from 453 MHz to 458 MHz while the 463 MHz to 468 MHz band for the signal transmission from to the cell sites to the motile terminals. Due to the rapid success of the initial NMT system and the limited capacity of the original design, another pair of frequency bands, i.e., 890-915 MHz and 935-960 MHz, was allocated in 1986 to support the expansion of user scale. The cell sizes in an NMT network are flexible, ranging from 2 km to 30 km, enabling a scalable system that can start with large cells and later accommodate more and more subscribers by shrinking the cell size. To serve car phones, NMT utilized a transmission power of up to 15 W (NMT-450) and 6 W (NMT-900), while the power was lowered (up to 1 W) for personal handsets. The NMT-450 system uses a lower frequency and higher maximum transmitter power level for large cell coverage, while the NMT-900 system uses a higher frequency and a lower transmit power to increase system capacity.

NMT employed FDMA to multiplex the signals of different mobile users, where the whole frequency band was divided into a magnitude of narrowband channels with a bandwidth of 25 kHz. Although voice channels are analog modulating the speech signals through FM, the control signaling between the cell sites and the mobile station was digitized. The control channels adopt the fast frequency-shift keying (FFSK) modulation to realize a data rate of up to 1200 bps. NMT is the first cellular system with the feature of fully automatic switching (dialing) and supported the handoff among cells from the initial version. Due to a joint standard of Nordic countries, it has a demand for international roaming. Hence, it naturally became the first cellular system worldwide to implement international roaming. The NMT specifications were free and open, allowing many companies such as Nokia and Ericsson to produce network equipment and pushing the deployment cost down.

2.3.3 Total Access Communications System/TACS

TACS is an analog cellular system developed in the United Kingdom, as a derivative of the US AMPS system. Its major modifications include changes to the operating frequency band, channel bandwidth, and data signaling rates. In 1985, Vodafone opened the first TACS network in the UK to offer analog cellular voice service. It achieved great commercial success and more than 25 countries adopted Total Access Communications System (TACS) as their 1G cellular system after its introduction in the UK. The TACS standard has also been modified for use in Japan under the name of Japanese Total Access Communication System (JTACs), which has different frequency bands and different numbers of analog channels.

The TACS system was initially allocated a pair of 25 MHz spectrum at the 900 MHz band. Using FDD, 890 MHz to 915 MHz and 935 MHz to 960 MHz were assigned for the uplink and downlink transmission, respectively. In the 1990s, the GSM digital cellular system was introduced at these frequency bands. TACS adopted a narrower channel bandwidth of 25 kHz compared with the original 30 kHz channels in AMPS. Using freuqnecy-division multiple access (FDMA), the system offers 1000 duplex channels in the 900 MHz band. The narrower bandwidth results in a reduced data rate of 8 kbps in the control channels, compared to 10 kbps in AMPS. To meet the rapid growth of market demand, an extra 16 MHz spectrum was issued for the evolution of TACS named Extended TACS (ETACS) to have more channels available. The TACS system has also been modified to create the Narrowband TACS (NTACS) system. It halved the channel bandwidth from 25 kHz to 12.5 kHz and changed the in-band 8 kbps signaling on the voice channel to 100 bps subband digital signaling.

2.3.4 Mobile Cellular System/MCS

In 1979, Japan launched the world's first commercial cellular system named MCS, which was developed and operated by NTT. The initial standard, MCS-L1, was designed to operate in the 800 MHz frequency band. Using the FDD mode, the signal transmission from the mobile terminals to the cell sites was assigned to the frequency band from 870 MHz to 885 MHz while the 925 MHz to 940 MHz band for the signal transmission from to the cell sites to the motile terminals. MCS adopted analog modulation like FM over each narrow channel with a bandwidth of 25 kHz. The control channels with a data signaling rate of 300 bps simulcast from all cell sites in a service area, restricting the maximum capacity of the MCS-L1 system.

The second-generation analog cellular system, MCS-L2, was developed to enhance the system capacity (Watanabe & Imamura, 1989). It commenced in May 1988 by overlaying on the legacy MCS-L1 network and sharing the 2×15 MHz spectrum at the 800 MHz band. New technologies for increasing spectrum efficiencies such as Sectorization, flexible channel assignment, and radio channel interleaving were applied to MCS-L2. The channel bandwidth was reduced from 25 kHz to 12.5 kHz with 6.25 kHz interleaving. This doubles the system's capacity of MCS. The control channels can either be in-band

signaling at 100 bps or subband digital audio signaling at 150 bps. MCS-L2 mobile telephones adopted receiving diversity technique. While the cost and size of mobile telephones increase, the performance and coverage of the system are improved.

2.3.5 C-450

Radio Telephone Network C (abbreviated as C-Netz in German) was a 1G analog cellular system developed and operated in West Germany by Deutsche Telekom. It is based on the C-450 standard developed by Siemens, as the successor of previous pre-cellular systems (A-Netz and B-Netz). It was updated to the second-generation digital cellular systems named D-Netz and E-Netz, both based on GSM standards and operating in 900 MHz and 1800 MHz bands, respectively. In 1985, C-Netz was commercially available to the public. Due to problems with the B-Netz networks, early adoption of C-Netz was very high, especially in rural areas which had lacked B-Netz coverage. The C-Netz network covered West Germany and West Berlin, but following German reunification in 1990, it was rapidly expanded to the whole of Germany. By 1988, the number of C-Netz subscribers grew to nearly 100,000 and reached a peak user base of more than 1.5 million subscribers across Germany. It remained in operation until 2000 when it was finally phased out to make way for newer digital cellular technologies. The C-Netz system was allocated a pair of 4.44 MHz spectra at the 450 MHz band. The frequency band for the uplink and downlink transmission is 451.3 MHz to 455.74 MHz and 461.3 MHz to 465.74 MHz, respectively. The primary channel bandwidth was 20 kHz with 10 kHz interleaving, and it also has a narrowband mode of only 12 kHz. C-450 standard has also been used in the 1G analog cellular networks of South Africa in 1986 and Portugal in 1989.

2.3.6 Radiocom2000

Radiocom2000 was developed by France in the 1980s by a consortium of French companies, including Alcatel, CGE, and Thomson, as their 1G analog cellular standard. In 1986, France Telecom/Orange launched the first Radiocom2000 network to provide mobile telephone service mainly to the devices mounted on vehicles. The initial system operated in the 400 MHz frequency band, where the uplink transmission was assigned to 414.8 MHz to 418 MHz and the downlink transmission was over 424.8 MHz to 428 MHz. The frequency separation is smaller than other 1G analog cellular systems, only 10 MHz, instead of 40 MHz in AMPS and NMT. It adopted a narrower channel bandwidth of 12.5 kHz, over which analog speech signals were modulated by FM. Using FDMA, the initial system offers 256 duplex channels in the 400 MHz band. Like other 1G cellular systems, Radiocom2000 used digital modulation for signaling in the control channels, despite analog modulation in the voice channels, where FSK modulation with a rate of 1200 bps and Hagelbarger coding were employed. Spectrum allocation in Radiocom2000 was more flexible and can be dynamically assigned based on the growing demand for subscriptions. In particular, the Radiocom2000 system also operated in the 160 MHz and 200 MHz bands in the most populous areas centered on the capital Paris, as well as the 175 MHz band in Lyon and Marseille regions. To meet the demand for more capacity, the system also operated in the 900 MHz frequency band since 1990.

The main technological features of the 1G analog cellular standards are summarized in Table 2.1.

Table 2.1 1G Cellular Standards (Goldsmith, 2005)

Standard	AMPS	NMT	MCS	TACS	C-450	Radiocom2000
Country/region	USA	Scandinavia	Japan	UK	Germany	France
Launch Year	1983	1981	1979	1983	1985	1986
Downlink [MHz]	869-894	463-468	870-885	935-960	461.3-465.74	424.8-428
Uplink [MHz]	824-849	453-458	925-940	890-915	451.3-455.74	414.8-418
Bandwidth [kHz]	30	25	25	25	20	12.5
Num. of channels	832	180	600	1000	220	256
Multiple Access	freuqnecy-division multiple access					
Duplexing	frequency-division duplex					
Modulation	frequency modulation					

2.4 Key Technologies for 1G Analog Cellular

Through the study of Sect. 2.3, the readers may draw a conclusion: Although these 1G cellular standards were developed by different countries or regions, the same major technologies were adopted. That is because they represent state-of-the-art wireless communications technologies and electronic components available at that time. This section outlines the fundamental components of 1G analog cellular networks, including frequency reuse, cell splitting, sectorization, handover, freuqnecy-division multiple access, and frequency-division duplex, which empowered the success of 1G analog cellular system and paved the way for the development of more advanced second-generation digital cellular technologies.

2.4.1 Frequency Reuse

In early pre-cellular systems, a metropolitan area was serviced by a single radio site that covered the entire service area with a radius of tens of kilometers. To achieve this coverage, the base station's antenna was generally installed at a high elevation, and it transmitted radio signals with high power in the low-frequency band, which has low propagation loss and high penetration capability. All mobile subscribers within this wide coverage shared the allocated spectrum. Frequency reuse was only possible among remote cell sites separated by approximately 100 km, leading to discontinuous service coverage in-between two cities and a limited system capacity. With the increasing demand for mobile telephone services, the need for a capable, economical, and portable system fostered the advent of the cellular network. In 1947, William R. Young, an engineer at AT&T Bell Labs, introduced the cellular concept of hexagonal geometry for a wide coverage area. Douglas H. Ring, also at Bell Labs, expanded Young's initial concept and sketched out the basic design for a standard cellular network, which he published as a technical memorandum titled *Mobile Telephony—Wide-Area Coverage* in Bell Labs' internal journal on December 11, 1947 (Ring, 1947).

The most critical component behind cellular network design is *frequency reuse* (Donald, 1979). It is based on the fact that radio signals attenuate significantly as they propagate through free space, making it possible to reuse the same frequency spectrum across spatially separated locations with acceptable levels of co-channel interference. To implement frequency reuse, base stations with moderate power are deliberately distributed throughout the coverage area, with each base station covering a nearby zone called a cell. Each cell operates on a specific set of frequencies, which can be reused in other cells that are sufficiently far away to avoid co-channel interference. The exact frequency reuse scheme used in a cellular network will depend on a variety of factors, such as the number of cells in the network, the amount of available spectrum, the channel bandwidth, and the specific characteristics of the propagation environment. However, a widely used approach is to use a pattern of hexagonal cells, with each cell being surrounded by six neighboring cells. In this scheme, cells that are separated by at least two intervening cells can use the same frequency set, thereby thus enabling efficient utilization of the available spectrum.

As illustrated in Fig. 2.5, the available spectrum is divided into a number of N narrowband channels. To minimize co-channel interference, the same channel is generally not used in neighboring cells. The channels allocated to a cell are not necessarily continuous, and it is better to assign non-continuous channels to suppress adjacent-channel interference. Each cell gets n channels. The ratio N/n is termed as the *frequency reuse factor*, which is a parameter used to indicate the number of cells that can reuse the same set of frequencies. It is defined as the ratio of the total number of cells in a system to the number of cells that are using the same frequency set. For example, the reuse factor is 7 in a classical hexagonal layout, where the channels are divided into seven groups denoted by $\{f1, f2, f3, f4, f5, f6, f7\}$. In this case, every seven neighboring cells, named a cluster of cells, share the whole spectrum.

The frequency reuse factor is an important consideration in network planning, as it affects both the capacity and coverage of a cellular network. While a higher reuse factor can improve spectrum utilization, it raises the level of co-channel interference that can degrade network performance. Therefore, network designers must carefully strike a balance between the benefits and drawbacks of different frequency reuse factors to achieve optimal system performance. Depending on the geometry of the cellular arrangement and the interference avoidance pattern, the optimal reuse factor can be different. For instance, AMPS adopted a factor of 7, while GSM used a factor of 3. CDMA-based cellular systems, on the other hand, have the capability of offering universal frequency reuse with a factor of 1. Thanks to advanced interference suppression

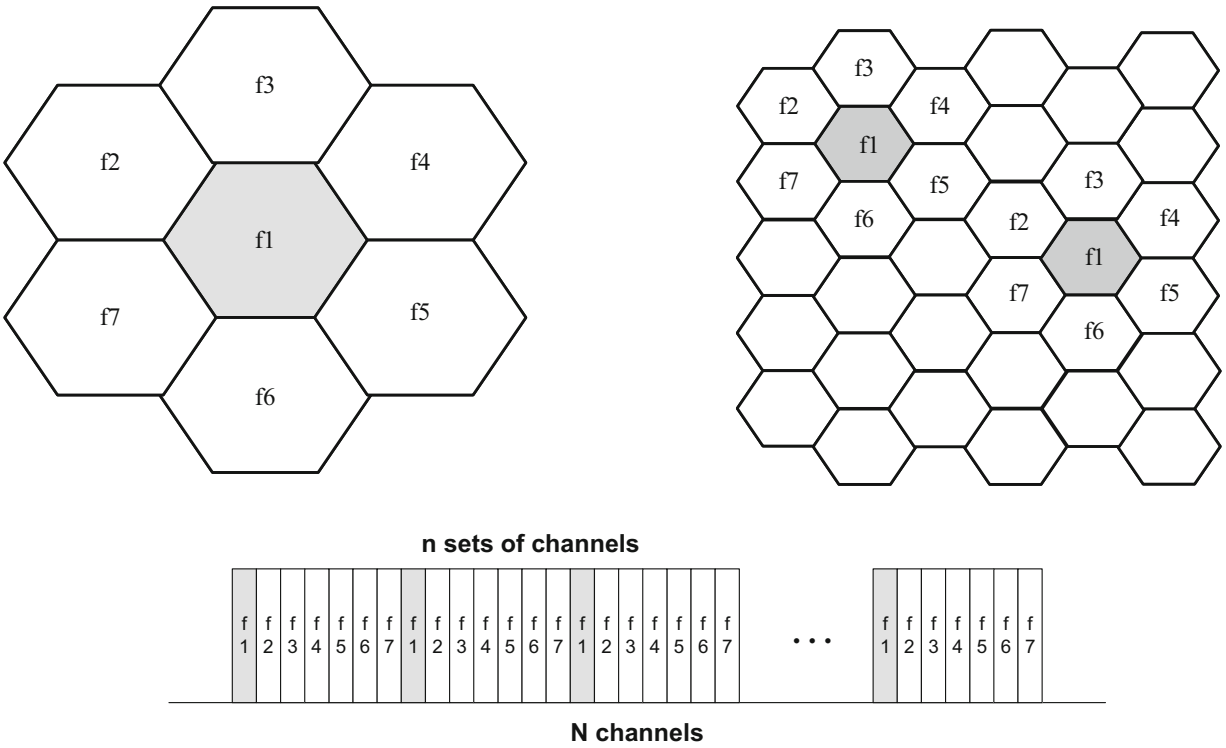


Fig. 2.5 Cellular network layout illustrating frequency reuse and cell splitting

techniques, the entire spectrum was reused in each cell. In addition to basic hexagonal geometry, researchers proposed more advanced frequency reuse schemes in recent years, including fractional frequency reuse and soft frequency reuse (Yang, 2014), allowing for different levels of frequency reuse in different parts of a cell.

2.4.2 Cell Splitting

Cell splitting is a critical technique used in cellular networks to ensure network scalability to handle the increasing demand for wireless services. The fundamental concept behind this technique is to partition a single large cell into several smaller cells, each with its own base station and a set of assigned frequency channels. In the process of splitting a cell, the power of the base station is decreased to cover a smaller area, and additional base stations are deployed to cover the original area. By reducing the size of each cell, the total number of available channels increases, which, in turn, enables more users to be supported within a given area. As a result, the reuse of the frequency spectrum within the same geographical area is enhanced, thereby augmenting the network capacity (Mishra, 2005).

During the initial phase of deploying a cellular network, it is efficient to cover an entire region or city with a few large cells, which are commonly referred to as macro-cells. The base station of a macro-cell is usually mounted on a high tower on the top of a tall building or a mountain, and it transmits with relatively high power. This arrangement is mainly driven by two factors: the high cost of the hardware and the low density of mobile subscriptions at the early deployment stage. As the number of users increases and begins to exceed the network capacity, obtaining additional bandwidth may appear to be a straightforward solution. Obtaining additional bandwidth may seem like the obvious solution. However, this is usually impractical due to the high cost and scarcity of licensed spectrum in some regions. As an alternative, cell splitting provides an economical approach to increase network capacity without acquiring additional spectrum. Figure 2.5 illustrates this approach, where a large cell is divided into a cluster of seven small cells, and the utilization of the frequency spectrum $\{f1, f2, f3, f4, f5, f6, f7\}$ increases sevenfold. With appropriate channel arrangements among cells, it is possible to have large and small cells coexisting within the same area. This enables the utilization of small cells in metropolitan regions with high traffic density while retaining large cells in the surrounding, less populated, and thus lower traffic level areas.

2.4.3 Sectorization

Sectorization is a technique employed in mobile communications to increase the capacity of a cellular system. Traditionally, a base station is equipped with an omnidirectional antenna where the signal power is radiated uniformly in all directions. Inspired by splitting a large cell into several small cells to improve spectrum efficiency, a cell is further divided into smaller areas called sectors, each with its own directional antenna. By using directional antennas, the same frequency can be reused in different sectors, allowing more users without building new sites or network infrastructure. In 1985, Philip T. Porter proposed the use of directional antennas at base stations, which could lower interference and enable a seven-cell reuse pattern (Porter, 1985).

The capacity of a cell can be tripled in theory by installing three directional antennas on a single cell site, with each antenna covering a 120°-degree sector. This arrangement is commonly referred to as a three-sector site, and in some cases, additional directional antennas can be added to create more sectors, such as six or even nine sectors. The application of sectorization can improve the overall system capacity and the quality of service experienced by users, particularly in congested urban areas with high call traffic. Furthermore, it can reduce interference levels and improve the signal-to-noise ratio, leading to better call quality, fewer dropped calls, and improved data throughput. It is worth noting that sectorization is more effective when the base station is tall with few surrounding obstacles, as inter-sector interference can occur in the presence of scattering and reflectors.

2.4.4 Handover

The use of frequency reuse and cell splitting allows a cellular network to offer ubiquitous coverage across a wide area, serve a large number of mobile users with limited spectrum allocation, and provide a scalable network that adapts to the growth of mobile traffic. Nevertheless, it raises a new problem—when a mobile user moves from one cell to another cell, the communication quality decreases or the connection is interrupted. To ensure uninterrupted service and guarantee the experience for mobile users, the connection needs to transition seamlessly between two cells through a process called handover (also known as handoff). The process of handover is regulated by several criteria, including the signal quality, the distance from the mobile station to the base station, and the network load of the target base station. The system designers and network operators must be carefully considered to ensure optimal network performance. The signal quality criterion is based on the strength of the received signal, which must meet a certain level of quality to be considered suitable for handover. The distance criterion considers the proximity of the mobile station to the new base station, as moving to a base station that is too far away may result in a weaker signal or network congestion. Finally, the network load criterion accounts for the current network traffic and determines if the new base station has the capacity to handle the additional load.

Typically, a cellular network conducts handover based on the signal strength. Each base station broadcasts a beacon signal at a consistent power level, which is distinguishable from the beacon signals of other base stations via the utilization of pseudo-random sequences. The operating carrier frequency, signal format, power level, and the definition of pseudo-random sequences are predetermined and commonly known by all mobile stations within the network. Periodically, a mobile station measures the strength of the beacon signal from the surrounding base stations. In the case of a mobile station situated at the center of a cell, the beacon signal from that cell is strong, while signals from other cells are weak. As a mobile station moves away from the center of this cell, the measured signal strength gradually diminishes. Once the signal strength falls below the predefined threshold required for acceptable performance, a handover procedure is initiated. When at the edge of a cell, a mobile station may receive several beacon signals with comparable strength. It selects the most appropriate cell with the strongest signal strength to access. The involved base stations and the controller assist the mobile station to release the occupied channel in the outgoing cell and tune to a new channel in the incoming cell.

The phenomenon of interest in handover, known as *the ping-pong effect*, is characterized by the repetitive switching of the mobile station between two adjacent base stations when the signal strength of each is similar. Frequent handovers due to the ping-pong effect result in dropped calls, reduced call quality, and lower data transfer rates. To mitigate this effect, network operators employ various techniques, such as setting appropriate handover thresholds, adjusting power levels of base stations, and implementing handover algorithms that leverage past handover history to make efficient decisions. Moreover, advanced handover methods, such as predictive handover, have been proposed to further enhance network performance. Predictive handover utilizes predictive algorithms to anticipate the need for handover before the signal strength falls below a threshold and searches for a new base station before the signal quality degrades below the threshold. Such an approach can reduce handover latency, ensure smoother transitions between base stations, and improve the quality of service provided to mobile users.

2.4.5 Frequency-Division Multiple Access/FDMA

In contrast to the unidirectional nature of radio and television broadcasting systems, mobile communications are two-way with dedicated messages for individual users. Such a multi-user system requires the allocation of resources to specific users, referred to as multiple access. Real-time applications such as voice or video communications require dedicated channels to ensure uninterrupted signal transmission. To achieve this, orthogonal channelization techniques such as frequency-division, time-division, space-division, or hybrid combinations are employed to create dedicated channels. On the other hand, delay-tolerant services, such as bursty data delivery, typically utilize non-orthogonal multiple access, also known as random access.

Due to its simple implementation and low complexity, different 1G systems such as AMPS, NMT, TACS, and C-450 selected FDMA as their multiple access approach. In FDMA, the system bandwidth is divided along the frequency axis into multiple narrowband channels, as shown in Fig. 2.6. Each channel is allocated to a different user and is used exclusively by that user for transmitting and receiving signals. This dedicated channel assignment ensures that there is no interference between users, thereby increasing the quality of voice calls. Since each narrowband channel suffers from frequency-flat fading, it does not require complex signal processing, making it a cost-effective solution for mobile communication systems. Another advantage of FDMA is its compatibility with legacy systems, making it an ideal choice for upgrading or enhancing existing networks.

Despite its advantages, there are also some disadvantages of FDMA. Suffering from the impairments, such as imperfect hardware, spectrum spreading due to the Doppler shift, and adjacent-channel spectral leakage, an FDMA channel has to use guard bands at both sides. It leads to the waste of spectral resources. For example, each AMPS user is assigned a 30 kHz channel, corresponding to 24 kHz for the FM signal transmission and 3 kHz guard bands on each side. This guard band wastes the available bandwidth, which limits the number of channels that can be allocated and reduces the capacity of the system. A terminal also needs to frequency-agile RF components that can tune to different channels. FDMA divides the frequency band into a fixed number of narrowband channels. This makes it requires careful frequency planning and management to ensure that the available frequency spectrum is allocated efficiently. It is less efficient than other multiple access methods, such as TDMA or CDMA in terms of capacity, and lacks flexibility in system design.

2.4.6 Frequency-Division Duplexing/FDD

A mobile communication system needs the ability to support two parties to communicate with each other simultaneously. Duplexing is a fundamental technique in mobile communications that allows for two-way communication over a single channel. There are two basic forms of duplexing techniques used in mobile communications: frequency-division duplex or

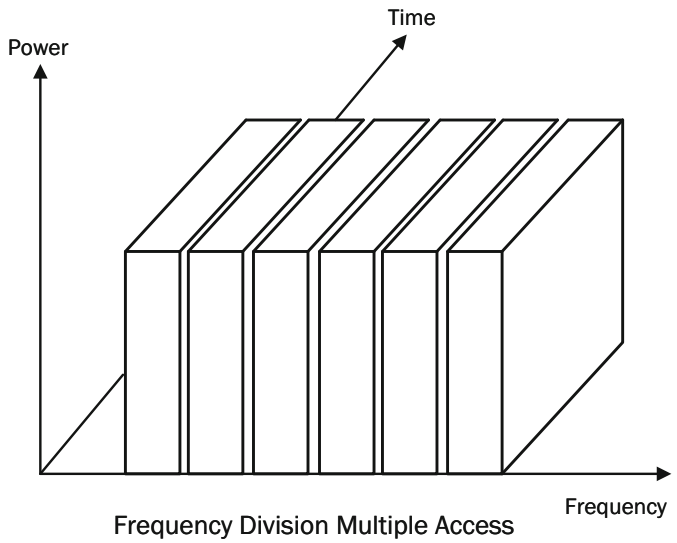


Fig. 2.6 Illustration of FDMA

FDD and time-division duplex (TDD), each with its own set of advantages and disadvantages, and they are used in different wireless systems depending on their requirements (Chan et al., 2006).

FDD employs two separate frequency bands for the uplink transmission (from a mobile terminal to a base station) and downlink transmission (from a base station to a mobile terminal). In FDD, the transmitter and receiver operate on a pair of frequencies, with a sufficient frequency separation between the uplink and downlink to minimize self-interference. FDD allows for simultaneous transmission and reception, making it ideal for real-time applications, such as phone calls or video conferences. FDD is commonly used in mobile communication systems from 1G to 5G cellular networks, as well as Wi-Fi and satellite communications.

In contrast, TDD uses a single frequency band for both downlink and uplink but separates the transmission and reception of data in time. In TDD, the base station and mobile terminal take turns transmitting and receiving signals on the same frequency channel. One unique advantage of TDD is that it can optimize the ratio between the downlink and uplink transmission by dynamically allocating time slots based on the amount of data traffic. Hence, it is particularly useful in data applications with an unbalanced amount of transmission and reception data. Furthermore, TDD is more flexible than FDD in terms of spectrum allocation, as it can exploit fragmented frequency bands instead of requiring a pair of frequency bands.

2.5 Summary

This chapter presented a comprehensive overview of the 1G systems that were developed based on the concept of cellular networking. We provided a brief summary of pre-cellular systems in order to contextualize the early evolution of mobile communications. The chapter then examined the various standards that comprise 1G, which were developed by different countries. Finally, we discussed key technologies that drove the evolution of 1G, seeking a complete understanding of the historical and technical foundations of this early evolution in cellular communications. The emergence of 1G cellular technologies revolutionized mobile communications, but its primitive design and technological limitations cannot support further growth of market demand. The advancements in digital technology in the 1980s paved the way for an upgrade from 1G analog cellular to 2G digital cellular systems, which offer a host of advantages, including greater capacity, higher quality of service, and improved security. In the next chapter, an overview of 2G digital cellular will be offered.

2.6 Exercises

1. In which year the concept of cellular networks has been proposed? By whom? Describe the main reasons that motivated the design of this novel networking paradigm.
2. The FCC had assigned a pair of frequency bands—824 MHz to 849 MHz and 869 MHz to 894 MHz—for the deployment of AMPS systems. Each analog channel has a bandwidth of 30 kHz. Calculate the capacity of an AMPS system.
3. The first-generation cellular systems are comprised of diverse standards developed from different countries or regions. How many 1G standards can you name? What are the common technical features of these standards?
4. Duplexing is a fundamental technique in mobile communications that allows for two-way communication over a single channel. Describe two basic forms of duplexing techniques used in mobile communications, and compare their pros and cons.
5. What are the technological pillars to implement cellular networking?
6. At the initial stage of deploying a cellular network, the number of mobile subscribers is small. It is reasonable to deploy a relatively small number of base stations from the financial perspective. When the scale of subscription raises, the network operator has to increase the system capacity. How can you upgrade the network smoothly? Which technology is involved?
7. Describe the fundamental idea of sectorization? What are its benefits?
8. Why does a cellular network need to apply a handover mechanism?
9. In cellular systems, *multiple access* is a frequently mentioned term. Why do we need multiple access? Describe the fundamentals of this technique.
10. Different 1G systems such as AMPS, NMT, TACS, and C-450 selected FDMA as their multiple access approach. What are the advantages of FDMA that were fit for 1G?

Evolution to Second-Generation (2G) Mobile Cellular Communications

3

3.1 From 1G Analog to 2G Digital Cellular

Similar to the earlier versions of any product or service, analog cellular systems were not initially referred to as "1G" until around 20 years later, when the term "3G" was coined to describe third-generation systems. While 1G marked the first leap of mobile cellular communications, the system was considered primitive and exhibited many deficiencies such as:

- Worse voice quality: 1G employed analog modulation that converts voice signals into electrical signals to be transmitted over narrowband frequency channels. It is vulnerable to interference, but channel coding and interference-resist technologies based on digital signal processing cannot be applied to avoid performance degradation. Moreover, analog signals were affected by the distance between the cell site and the mobile station, leading to unreliable voice quality.
- Limited system capacity: 1G usually adopted large cells with a high frequency reuse factor, where each cell only uses a small portion of the available frequency spectrum. Digital voice compression that enables more calls to be carried on a single cell site cannot be implemented in analog systems. Hence, 1G has a limited capacity and could only support a small number of users simultaneously.
- Limited coverage: During the time of 1G, the cost of deploying a cellular network was high due to the small scale of production, which meant that many areas did not have the necessary infrastructure to support ubiquitous coverage. In addition, the analog technology limits the transmit power and propagation range of 1G systems, making it particularly challenging to provide coverage in remote areas or locations with a low population density.
- No security protection: Eavesdropping on calls was common during the 1G era, as analog signals were vulnerable to interception but no encryption technology was available. In addition, the system designers did not fully take into account security concerns, exposing the network to various forms of attacks. For instance, the identification information of a mobile station could be easily acquired by a fake base station and used illegally in other locations.
- No Data Services: 1G cellular networks were primarily designed for voice communications and did not support any data services. Users could not send text messages, access the Internet, or send emails through their mobile devices.
- Limited international roaming: During the 1G era, mobile phones were still not widely used, and international travel was not so prevalent. As a result, the need for international roaming was not as high as it is today. Moreover, the 1G standards were developed independently by different countries. The lack of standardization led to different frequency bands and incompatible air interfaces, which are challenging to support roaming between different networks. This situation prompted the introduction of 2G networks, which were designed with standardized protocols and frequency bands that facilitated international roaming.
- Poor handover reliability: The handover process in 1G networks was based on the measurement of signal strength, which was sometimes inaccurate due to the vulnerability of analog signals to interference. Furthermore, the early 1G networks suffered from limited network capacity and coverage, posing challenges for ensuring seamless handovers as mobile users transitioned between different cells. As a result, the handover process was often subject to delays and errors, degrading the overall quality of service.
- Less affordability: The cell phone was quite expensive at that time, and the subscription fee for mobile call service was also very high, blocking its massive use in the public.

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5_3

The transition to 2G cellular systems was driven by the need for enhanced capacity while providing high-quality, cost-effective, and secure services. The mobile industry initiated the development of new-generation technologies in the early 1980s, and it gradually replaced analog systems in the 1990s. 2G is a generation of digital cellular systems empowered by digital technology. Digital technology offers a number of advantages over analog technology. Digitized voice signals can be compressed, which supports more users simultaneously in the same amount of spectrum, increasing the capacity of a cellular network. Digital signals are also less susceptible to interference and are easier to protect through channel coding, leading to an improved quality of service. Digital technology enables the encryption of voice and data transmissions, promoting increased security and privacy for users. The transition from 1G to 2G was a significant milestone in the history of cellular communications. It marked the beginning of a new era of mobile communications, where digital technology played a central role in enabling greater capacity, improved quality of service, and the development of new data services. The transition to 2G paved the way for further advancements in mobile communications.

Digital components are more efficient, lightweight, compact, affordable, and power-efficient than their analog counterparts, improving the compactness and affordability of mobile phones with prolonged battery life. The evolution from 1G analog to 2G digital technology has facilitated the innovations of mobile terminals, which have become smaller, lighter, cheaper, and more power-efficient, thus leading to the widespread adoption of mobile phones as an integral part of contemporary life. Among the companies that have leveraged this trend, Nokia stands out for its shrewd recognition of the desire for personalized mobile devices. In 1994, Nokia released the first cell phone to use the iconic ringtone—Nokia 2110, the first cell phone allowing the user to change the phone's covers to reflect their mood or style—Nokia 5110, and the first to feature the mobile game Snake—Nokia 6110. Consequently, Nokia attained a significant market share in the global cell phone market. By 2002, the transition to digital cellular networks was complete, and the number of mobile subscribers exceeded that of fixed-line telephone subscribers for the first time, thus cementing cellular networks as the dominant communication service technology.

Digitization *The transition from the first-generation to the second-generation cellular system was empowered by digital technology. A digital system can achieve a higher capacity than an analog system since digital communications can apply more spectral-efficient digital modulation and more efficient multiple access techniques. Digitization facilitates the compression of voice signals, the encryption of information against eavesdropping, and the support for data services. In addition, digital components are more powerful, more lightweight, smaller, cheaper, and more power-efficient than analog components.*

3.2 2G Digital Cellular Standards

2G is comprised of a variety of standards that were developed by different countries and organizations to cater to specific market demands. The most widely used 2G standard was the Global System for Mobile Communications (GSM), which was developed by a consortium of European telecommunications companies. GSM employed TDMA and digital modulation techniques to transmit voice and data signals, thereby enabling efficient use of the radio spectrum. GSM quickly became the dominant 2G standard worldwide, with numerous nations adopting it for their cellular networks. To support data services, GSM was smoothly upgraded to High Speed Circuit Switched Data (HSCSD), General Packet Radio Service (GPRS), and subsequently to Enhanced Data Rates for GSM Evolution (EDGE). In the United States, the development of the digital cellular system fell into a raging debate on the selection of spectrum sharing techniques between TDMA and Code-Division Multiple Access (CDMA). The outcome of this debate led to the creation of two incompatible systems: Interim Standard-54 (IS-54) and its evolution IS-136 versus IS-95. IS-54 and IS-136 constituted the Digital Advanced Mobile Phone System (D-AMPS), the digital upgrade to AMPS. On the other hand, Qualcomm developed IS-95, a CDMA-based standard used primarily in North America and parts of Asia. In Japan, the Personal Digital Cellular (PDC) standard was developed by NTT DoCoMo. PDC was designed to be a digital cellular system that was compatible with the existing analog cellular system in Japan.

3.2.1 Global System for Mobile communications/GSM

During the 1980s, Europe developed incompatible 1G standards including NMT in Scandinavia countries, TACS in the UK, C-450 in Germany, and Radiocom2000 in France. Despite their success in commercial use, these European analog cellular systems could not compete with AMPS, which enjoyed the backing of a large, unified mobile market in the United States, and was thus recognized as the most successful 1G analog standard. The incompatibility of these European systems made it difficult for travelers to maintain continuous communication service with a single mobile phone across European countries. It motivated the necessity for a uniform European standard and unified frequency allocation throughout Europe.

In response, the *European Conference of Postal and Telecommunications Administrations* (CEPT) established a working group in 1982 called the Groupe Special Mobile (the initial meaning of GSM) to coordinate the development work. Discussions were held between 1982 and 1985 in the GSM group to deliberate the selection between analog and digital technology. After several field trials, a digital cellular system based on narrowband TDMA was ultimately selected (Mouly & Pautet, 1995).

The following objectives for this pan-European cellular system were defined:

- Create a unified standard for a large market area that could be adopted by multiple operators across different countries. This approach has the potential to reduce the cost of network equipment and cell phones by taking advantage of the economics of scale, which arises from sharing the initial investment for research and development among a large number of customers.
- Support international roaming, allowing users to use their phones in different European countries without having to change their phone number or SIM card. Enable connecting a voice call to the GSM user regardless of which country the user is located in and which operator network the phone is currently using.
- Optimize the utilization of the spectrum allocated to GSM channels, thereby enabling a greater capacity for handling a larger number of users per cell.
- Offer superior voice services compared to existing analog cellular systems, with an emphasis on providing high-quality and secure services.
- Support data transmission and messaging services that would be compatible with the fixed Public Switched Telephone Network (PSTN) network, especially with the Integrated Services Digital Network (ISDN), which emerged parallel to GSM standardization in the 1980s.

In 1984, France and Germany signed a collaborative development agreement which was later joined by Italy and the UK in 1986. During the same year, the European Commission recommended reserving the 900 MHz spectrum band for the use of the GSM network, where 890 MHz to 915 MHz and 935 MHz to 960 MHz were assigned for the uplink and downlink transmission, respectively. In 1987, a memorandum of understanding was signed in Copenhagen by representatives from 13 European countries with the goal of creating a common cellular system throughout Europe. This agreement led to the application of European Union regulations making GSM a mandatory standard. This choice to create a unified standard for the continent eventually led to the creation of an open, standard-based network that was larger than its US counterpart. In 1988, the CEPT formed a new standardization forum, *European Telecommunications Standards Institute* (ETSI), and the responsibility for specifying the GSM standard was transferred to ETSI in the subsequent year (Rahnema, 1993). GSM standards have three major phases of evolution:

- Phase 1 recommendations for the GSM standard were released in 1990. Along with the basic voice service, it included specifications for Short Message Service (SMS) and supports to connect with the ISDN for various data services with a rate of 9.6 kbps. In addition, it introduced a number of basic supplementary services such as call forwarding and network roaming, which are now commonly associated with cellular networks. While the air interface and SMS were new features for GSM, its core network relied on legacy circuit-switched technologies like digital exchanges and the Signalling System No. 7 (SS7) protocol stack. GSM expanded SS7 protocols with mobility and radio resource management specifically for cellular networks. This allowed vendors to support GSM using their off-the-shelf products with just additional software packages.
- Phase 2 recommendations for the GSM standard were finalized in 1995. It introduced several new features and enhancements to GSM Phase 1 like conference calls, call waiting, call hold, and caller identification presentation, which displays the caller's number on the GSM phone of the callee. Furthermore, the specifications defined how GSM could be utilized for transferring data and telefaxes.

- Phase 2+ recommendations for the GSM standard were frozen in 1997. The primary objective of Phase 2+ was to improve data rates by introducing High Speed Circuit-Switched Data (HSCSD) on the circuit-switched network in release 96 and incorporating packet-switched data support on GSM networks via the GPRS in release 97 of the standard. In addition to SMS, Phase 2+ enabled multimedia messaging service (MMS), allowing users to send messages with multimedia content, such as images, videos, and other media. Location-based services, which allowed mobile operators to offer services like location-based advertising and emergency services, were introduced. Furthermore, the Phase 2+ standard included improved security measures through the implementation of advanced authentication and encryption techniques, thus ensuring user privacy and security.

Currently, the GSM standards are still maintained by the 3GPP standardization forum. In 3GPP, the GSM radio access network is referred to with a new acronym, GSM EDGE Radio Access Network (GERAN), including both the air interfaces of GSM and EDGE. The commercial launch of the first operational GSM network occurred on July 1, 1991, in Finland, operated by Radiolinja with equipment from Nokia and Siemens. It was long believed that the world's first GSM call was made on this day by Harri Holkeri, the former Finnish Prime Minister, who connected with Kaarina Suonio, the deputy mayor of Tampere. However, in 2021, a former Nokia engineer Pekka Lonka revealed that he had conducted a test call a few hours before Holkeri's call, as the actual first GSM call. A variant of the GSM standard operating on a higher frequency band was standardized within ETSI and received approval in February 1991. The first 1800 MHz network became operational in the UK by 1993, known as Digital Cellular System 1800MHz (DCS-1800) (Potter, 1992), where 1710 MHz to 1785 MHz and 1805 MHz to 1880 MHz were assigned for the uplink and downlink transmission, respectively. The same year also marked the deployment of the first GSM network outside of Europe built by Telstra, and the first practical handheld GSM cell phone became available in the market. In 1995, data and SMS messaging services were launched commercially.

The GSM standard achieved rapid acceptance worldwide and emerged as the leading 2G digital cellular standard (Vriendt et al., 2002). It experienced outstanding commercial success, with a global market share surpassing 90%. By early 2004, more than 1 billion population in more than 200 countries and territories enjoyed their mobile telephony services thanks to GSM.

3.2.2 Digital Advanced Mobile Phone System/D-AMPS

Due to its economy of scale achieved by the backing of a large, unified mobile market in the United States, the AMPS standard achieved a dominant position over sporadic and competitive European standards in the era of 1G. However, the United States did not continue the same success in the second round of cellular development. A heated debate ensued over the selection of spectrum sharing techniques between TDMA and CDMA for second-generation digital cellular technology, resulting in two incompatible systems: IS-54 (and its subsequent evolution IS-136) based on TDMA and IS-95, a CDMA-based standard developed by Qualcomm.

IS-54 and IS-136 constituted the D-AMPS, which was the digital evolution of the AMPS standard in the USA. IS-54 was backward compatible, inheriting the basic architecture and signaling protocols from its predecessor. It was deployed in the same frequency bands of AMPS, i.e., 869-894 MHz for the downlink and 824-849 MHz for the uplink. Dual-mode cell phones allowed users to continue using their analog AMPS services while taking advantage of the newer digital D-AMPS system, particularly in areas where D-AMPS coverage was limited. This deliberate design allowed for a smooth transition from an analog cellular system to a digital cellular system. Each 30 kHz channel that carries only a single voice user in analog AMPS systems was divided into three time slots using TDMA, with each slot carrying data for a different user. With digital compression of voice data, each channel is enhanced to accommodate three users, resulting in a tripled system capacity. IS-54 won over Motorola's Narrowband AMPS or N-AMPS, an analog scheme that increased capacity, by cutting down voice channels from 30 kHz to 10 kHz. IS-54 also supports SMS and other supplementary services like call waiting and caller identification presentation, which displays the caller's number on the GSM phone of the callee.

The specification of IS-54 was completed in 1992 and was deployed in the USA and Canada ever since its first commercial launch in 1993 by Ameritech in the Chicago area. It was enhanced over time, and these enhancements evolved into the IS-136 standard (Sollenberger et al., 1999). IS-136 introduced several new features to the original IS-54 standard, including circuit-switched data, text messaging, and the support of operating in 1900 MHz. IS-136 opened the possibility for an all-digital TDMA system instead of the dual-mode operation adopted by its predecessor. However, D-AMPS faced competition from other digital cellular standards like GSM and CDMA, which offered even greater capacity and features like global roaming.

As a result, D-AMPS saw a decline in usage in the late 1990s and early 2000s, and it was eventually phased out in favor of newer digital cellular technologies like 3G.

3.2.3 Interim Standard 95/IS-95

The development of the IS-95 standard began in the late 1980s by Qualcomm, aiming to create a digital cellular system that could support a large number of users and provide better call quality compared to analog cellular systems such as AMPS. Qualcomm's research team, led by Dr. Irwin Jacobs, developed a technique called CDMA, which separates each user's communication signals from one another using a spread-spectrum technique, allowing multiple users to communicate simultaneously within the same frequency channel. This enabled some unique technical advantages over TDMA for cellular systems, e.g., higher system capacity and simple frequency planning due to universal frequency reuse, high quality of service during soft handover, no hard limit on the number of users (soft capacity), the ability to exploit voice activity to reduce the aggregated interference automatically, and improved robustness using noise-like spread-spectrum signals. In contrast to previous narrowband mobile communications, it spreads information bits over a wideband channel using the direct-sequence spread-spectrum technique. It used a 1.25 MHz bandwidth, which allowed for 64 simultaneous voice calls per channel. The CDMA system required complicated air interfaces and communication protocols, e.g., the Rake receiver was adopted to mitigate the effect of multi-path transmission. Due to multi-user interference and inter-cell interference, the system performance depended heavily on accurate power control, especially in the uplink, to compensate for the near-far effect. A power control bit was transmitted 800 times per second on the forward link to instruct a mobile station to adjust its transmit power with a granularity of 1 dB.

In 1989, Qualcomm successfully demonstrated the world's first CDMA-based cellular system, and the initial specifications of the system were finalized in 1993. Telecommunications Industry Association (TIA) and Electronic Industries Alliance (EIA) of the United States approved it as a digital standard in 1995, hence named Interim Standard 1995 (IS-95) or IS-95A. In October 1995, Hutchison Telephone launched the world's first commercial CDMA cellular network in Hong Kong, under the name of cdmaOne. Due to its ability to provide better call quality, higher capacity, and improved security compared to the previous analog systems, IS-95 quickly gained popularity and was adopted by network operators in the United States, Asia, and Europe. There was much debate about the relative merits of the IS-54 and IS-95 standards throughout the early 1990s, claiming that IS-95 could achieve 20 times the capacity of AMPS, whereas IS-54 could only achieve three times this capacity. In the end, both systems turned out to achieve approximately the same capacity increase over AMPS (Goldsmith, 2005).

IS-95 systems offered circuit-mode and packet-mode data services at a data rate of up to 14.4 kbps. However, the explosive growth of the mobile Internet imposed an increasing demand for higher capacity, advanced multimedia services, and higher data rates. The evolution of the IS-95 standard to higher data rates and more advanced services occurs in two steps. The first step, IS-95B, is an enhancement to the IS-95 standard and offers the highest possible performance without breaking current IS-95 air interface design characteristics, thereby maintaining strict compatibility with existing base station hardware. IS-95B supports a high data rate of 64 kbps in both directions, and a new burst mode packet data service is defined to allow better interference management and capacity utilization. The second evolution step, CDMA2000, provides next-generation capacity, data rates, and services. The CDMA2000 system includes a greatly enhanced air interface supporting CDMA over wider bandwidths for improved capacity and higher data rates while also maintaining backward compatibility with IS-95 mobile devices. The CDMA2000 system also includes a sophisticated MAC feature to effectively support very high data rate services (up to 2 Mbps) and multiple concurrent data and voice services (Knisely et al., 1998). Despite being phased out by newer technologies, IS-95 remains an important part of the history of cellular communications and contributed to the development of modern wireless technology.

3.2.4 Personal Digital Cellular/PDC

The digital cellular standard known as PDC was independently developed by Japan and was exclusively deployed within the country. It was developed in the late 1980s and early 1990s by NTT and several other Japanese telecommunications companies as a successor to the analog cellular system. PDC, similar to D-AMPS and GSM, adopted TDMA as the multiple access technique. To maintain compatibility with Japanese analog systems, PDC selected a signal bandwidth of 25 kHz for voice channels, which were divided into three time slots for full rate (11.2 kbps) or six time slots for half-rate (5.6 kbps) voice

Table 3.1 Comparison of 2G Cellular Standards

	GSM	D-AMPS	PDC	IS-95
Launch Year	1991	1993	1993	1995
Downlink Band [MHz]	935–960	869–894	940–960, 1477–1501	869–894
Uplink Band [MHz]	890–915	824–849	810–830, 1429–1453	824–849
Bandwidth [kHz]	200	30	25	1250
System Capacity	1000	2500	3000	$\sim$ 2500 (soft)
Multiple Access	TDMA			CDMA
Receiver	Equalizer			RAKE
Duplexing	FDD			
Modulation	GMSK	$\pi/4$-DPSK	$\pi/4$-DPSK	BPSK/QPSK
Speech Rate [kbps]	13	7.95	11.2(full)/5.6(half)	1.2 $\sim$ 9.6 (variable)

codecs. The Research and Development Center for Radio System (RCR), now known as the Association of Radio Industries and Businesses (ARIB), finalized the specifications in April 1991. NTT DoCoMo launched its digital service in March 1993 using network equipment manufactured by NEC, Motorola, and Ericsson. After a peak of nearly 80 million subscribers, it was slowly phased out in favor of 3G technologies and was shut down on April 1, 2012. The PDC network offered mobile voice services (full- and half-rate), supplementary services (call waiting, voice mail, three-way calling, call forwarding, etc.), circuit-switched data service (up to 9.6 kbps), and packet-switched data service (up to 28.8 kbps).

Despite Japan's isolation from the rest of the world, its 2G network gave rise to an innovative technology known as i-mode. It was a mobile Internet service developed by Japan's mobile operator NTT DoCoMo in 1999. It was one of the earliest mobile Internet services in the world and was highly successful in Japan, with over 50 million users at its peak. The i-mode service allowed users to access a variety of Internet-based content and services from their mobile devices, including email, news, weather forecasts, entertainment, and e-commerce. One of the key features of i-mode was its use of a specially designed programming language called Compact HyperText Markup Language (HTML), which enabled content to be optimized for display on small screens of mobile devices with limited processing power. It paved the way for the development of other mobile Internet services and helped establish the importance of mobile devices in accessing online content. The technology developed for i-mode also influenced the development of future mobile Internet standards such as Wireless Application Protocol (WAP) and HTML5.

Table 3.1 provides a comprehensive comparison among different 2G digital standards.

3.3 2.5G Cellular Standards

With the proliferation of Internet services and mobile devices in the late 1990s, the demand for mobile data services increased rapidly. 2G cellular technology was primarily designed for optimizing voice communication, with limited support for data services. The development of 2.5G cellular systems, as an enhancement of 2G, was motivated by the increasing demand for faster data transmission speeds and the need for novel mobile data applications. Specifically, the development of 2.5G aimed to achieve the following objectives:

- Faster Data Transmission Speeds: 1G analog cellular systems provided only basic voice communication services. 2G introduced digital technology and allowed for the transmission of data along with voice, but it was limited in terms of data transmission speeds. The major motivation behind developing 2.5G was to offer a high data rate for mobile users.
- Support for Mobile Data Applications: With the increasing popularity of mobile phones, users wanted to use their phones for more than just voice communication. They wanted to use their phones for Internet browsing, email, and multimedia messaging. Previous circuit-switched networks were optimized for voice applications but did not support data services well. 2.5G systems introduced packet-switched infrastructure to support data transmission in a more efficient way.
- Transition to 3G: Another motivation behind the development of 2.5G was to serve as a transitional phase between 2G and 3G cellular systems. 2.5G is backward compatible with 2G, allowing for incremental improvements in data transmission speeds over legacy 2G networks, before the availability of high speed data services enabled by revolutionary 3G technologies.

3.3.1 High Speed Circuit-Switched Data/HSCSD

In addition to improved security due to digital encryption and significantly increased system capacity over their predecessors, another milestone progress of 2G was the introduction of data service into the mobile network. In 1992, SMS, with a data rate of 9.6 kbps over circuit-switched architecture, was born. Neil Papworth, a 22-year-old software engineer working at Vodafone, sent the world's first text message on December 3, 1992 when he typed "Merry Christmas" from a computer to Richard Jarvis on an Orbitel 901 handset. With the phenomenal success of SMS and the rising demand for accessing the Internet via mobile phones and laptop computers, the demand for high-rate data services boomed.

The circuit-switched connection provided by GSM is inadequate for sophisticated web browsing and the transfer of large files. GSM was primarily developed for voice telephony and has a channel spacing of 200 kHz. Although it offers a subset of bearer services from Integrated Services Digital Network (ISDN), allowing circuit-switched data connections of up to 9.6 kbps, this was only suitable for the time of the original system design when it compared favorably to fixed connection data rates. However, fixed connection data rates have since increased significantly, while the GSM channel structure and modulation technique are unable to support faster rates. As a result, the HSCSD service was introduced in the GSM Phase 2+ to provide faster data rates. As early as the 3GPP Release 96 of the GSM specifications, it was recognized that combining two or more channels into a single, grouped, the circuit-switched connection could lead to a significant increase in speed. Up to four channels can be merged according to the 3GPP TS 22.034 requirements and 3GPP TS 23.034 architecture, resulting in data rates of 57.6 kbps.

However, due to the circuit-switched connection tying up all four channels, the network operator was likely to charge a considerably higher rate for HSCSDcalls than for simple, single-channel calls. Therefore, the service never gained much popularity. HSCSD has since been largely replaced by more versatile, higher data rate, and more cost-effective alternatives such as GPRS and EDGE.

3.3.2 General Packet Radio Service/GPRS

Circuit-switched networks establish a dedicated physical or virtual connection between the sender and the receiver for the duration of the communication session. Prior to transmission, the connection must be established, and once the communication session is complete, it must be terminated to release resources. The bandwidth of a circuit-switched connection is fixed, and enough resources are reserved for the duration of the connection, guaranteeing high quality, low latency, and reliability for communication services. Data traffic is known for its high dynamism, with web browsing being a prime example. While users browse the web, they need short periods of high-rate downloads interspersed with long idle periods during which they digest the content. Additionally, web traffic is characterized by its asymmetric nature, as only small Hypertext Transfer Protocol (HTTP) queries are transmitted upstream, while large responses arrive downstream. Circuit-switched transmission is inefficient in such scenarios, as it suffers from long setup time, wastes network resources during idle periods, and slows down high-rate data bursts that exceed the constant rate of circuit-switched connections. By dividing burst data into small packets and transmitting them on a best-effort basis, packet-switched networks can handle dynamic and asymmetric traffic patterns effectively, ensuring the timely delivery of high-rate data bursts while efficiently using network resources.

In response to the earlier Cellular Digital Packet Data (CDPD), overlaying the AMPS system to provide a rate of 19.2 kbps, and Japanese i-mode services, ETSI standardized an enhancement of GSM called General Packet Radio Service. The Cellular Packet Radio (CELLPAC) protocol that introduced packet switching in GSM was the root for the specification of GPRS starting from 1993 (Walke, 2003). Through packet-switched networks, GPRS enables always-on connectivity, allowing users to stay connected to the Internet without having to establish a new connection each time they want to use it. Compared to the long setup time of circuit-switched connections, GPRS can be activated quickly by dynamically reserving one or multiple GSM time slots for the data transmission and later releasing them when there is nothing to be sent. It is also able to allocate different numbers of time slots to uplink versus downlink for asymmetric data transmission.

In June 2000, British Telecom Cellnet launched the world's first commercial GPRS network in the United Kingdom. GPRS is an overlaying packet-switched data network on the circuit-switched GSM network. Relying on the legacy air interface, an operator only needs to install some network nodes to upgrade a voice-only GSM network to a voice-plus-data GPRS network. Base station controllers separate the data and voice traffic and direct the data to GPRS support nodes connected to the data network. Operating in a best-effort style, GPRS typically reached a data rate of 40 kbps in the downlink and 14 kbps in the uplink by aggregating multiple time slots into one bearer. Enhancement in later specifications can theoretically achieve a peak rate of 171.2 kbps by aggregating eight time slots at the same time for a single user.

A generic GSM system is composed of four subsystems interconnected through interfaces (Rahnema, 1993), i.e.,

- Mobile Station Subsystem (MSS) is comprised of mobile equipment that provides support for voice calls, SMS, and low-speed data access. Unambiguous identification of a mobile station is ensured by assigning a unique serial number known as the International Mobile Equipment Identity (IMEI). In addition, the Subscriber Identity Module (SIM) card stores the International Mobile Subscriber Identity (IMSI), which serves to identify the subscriber, as well as a secret key for authentication and other relevant subscriber information. The independence of the IMEI and IMSI allows for the support of user mobility.
- Base Station Subsystem (BSS) is composed of BTS and Base Station Controller (BSC). The former comprises antennas for transmitting and receiving electromagnetic waves, transceivers for generating and detecting radio signals, and equipment for encryption and decryption. The latter is responsible for managing radio resources for one or more BTSs and controlling user admission, channel setup, inter-BTS handover, and frequency hopping.
- Network and Switching Subsystem (NSS) is to perform the switching function and establish connections between the serving mobile stations, other mobile users, and fixed telephony users. Additionally, it provides functionalities such as authentication, registration, location updating, handover, and call routing. Mobile Switching Center (MSC) serves as the central component of the NSS, aided by a set of databases, including Home Location Register (HLR) that holds the administrative details of local subscribers, Visitor Location Register (VLR) that stores information of visiting subscribers, Authentication Center (AuC) that stores the secret key for authentication and encryption, and Equipment Identity Register (EIR) maintaining a catalog of all valid terminals.
- Operation and Support Subsystem (OSS) enables network operators to effectively monitor and manage the system. In particular, Operations and Mainenance Center (OMC) is the equipment that connects to different components in the switching system and the BSC. OMC is in charge of the functions such as administration operation (subscription, termination, charging, and statistics), security management, network configuration, performance monitoring, and maintenance.

The specifications for GPRS were provided in Release 97 of the GSM standard (3GPP TS 23.060, 1999). It was implemented by overlaying a packet-switched sub-network on the existing circuit-switched network, as illustrated in Fig. 3.1. GSM networks can be smoothly upgraded to support data services by updating software and installing new network equipment including Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN).

SGSN performs a similar function to that of MSC in the circuit-switched network for voice traffic. It serves the mobile stations via the base stations and communicates with GGSN to provide access to external networks. Its functions include:

- Mobility management: When a mobile station attaches to the packet-switched network, the SGSN generates mobility management information according to the current location of mobile stations. The SGSN tracks the movement of a registered mobile station and forwards the incoming packets to the approximate address.
- Session management: The SGSN manages the initiating, maintaining, and terminating of real-time or non-real-time data sessions and provides mechanisms to guarantee the required quality of services for a variety of data services.
- Switching: The SGSN forwards incoming and outgoing data packets from the BSC and GGSN. It also communicates with other areas of the network, e.g., MSC and other circuit-switched areas, to get necessary management information.
- Charging: The SGSN is also responsible for charging and statistics collection by monitoring the flow of user data across the GPRS network. The SGSN generates the record of call details for the charging entities.

GGSN is responsible for the interworking of external packet data networks, including the Internet and X.25 networks. It is connected to SGSN through the GPRS backbone network and functions as a router from the perspective of external networks. GGSN is equipped with gateway features that enable it to publish subscriber addresses, map addresses, route and tunnel packets, screen messages, and count packets. At the GGSN, GPRS packets originating from SGSNs are converted to the appropriate data format before being forwarded to external data networks. Incoming data packets are converted to the appropriate format at the GGSN and forwarded to the SGSN associated with the target mobile station (Lin et al., 2001).

3.3.3 Enhanced Data Rates for GSM Evolution/EDGE

On the one hand, GPRS was found to exhibit certain limitations, such as low practical data rates that were significantly lower than the theoretical values. On the other hand, mobile network operators who were unable to obtain a 3G license were in need of a further-enhanced GPRS standard that could offer data services at speeds comparable to those available on

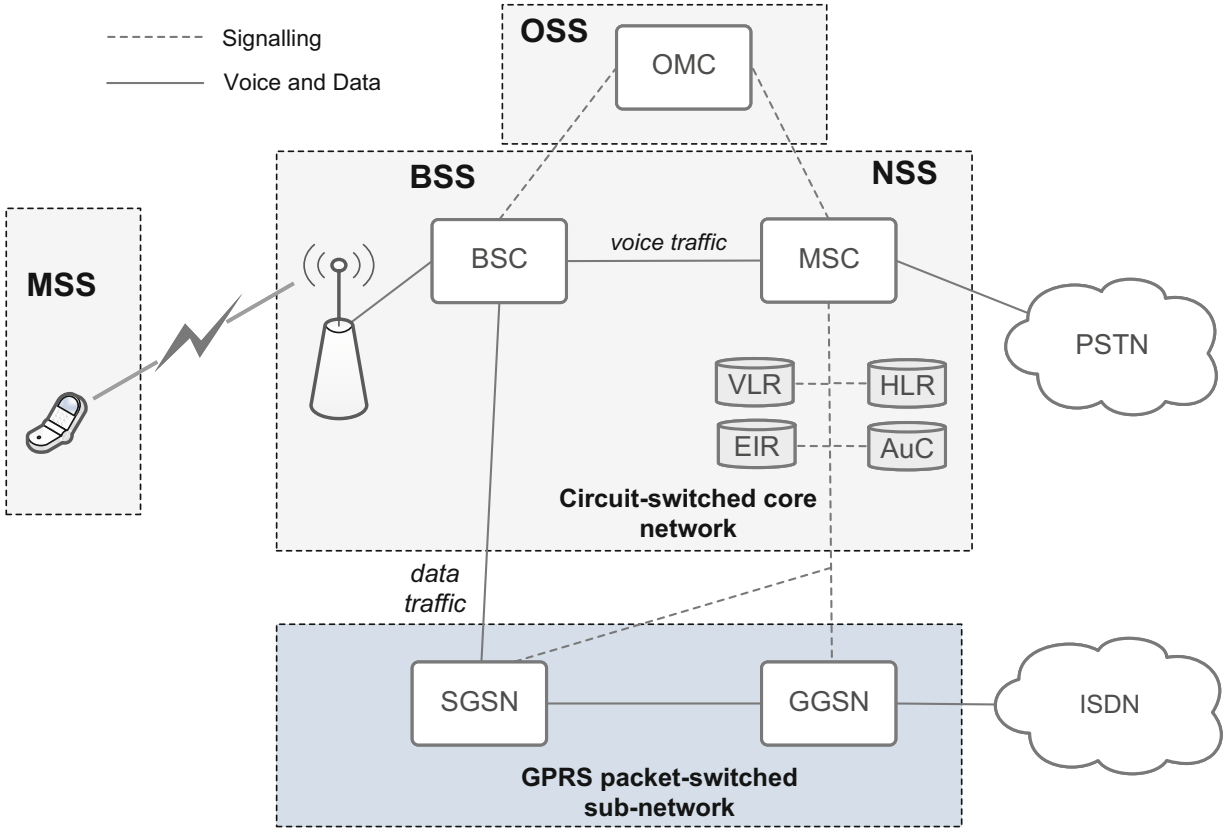


Fig. 3.1 The architecture of the GSM system with a circuit-switched core network, overlapped with a GPRS packet-switched sub-network

initial 3G networks. To address this need, as its name suggests, EDGE (Enhanced Data Rates for GSM Evolution) was first developed by the ETSI in 1997 as an enhancement of GPRS, aiming at providing faster data rates. EDGE comprises a set of specifications for both circuit-switched and packet-switched networks. Apart from providing improved data rates, EDGE is transparent to the service offering at the upper layers, which allows it to be applied on top of HSCSD, with the name of Enhanced Circuit Switched Data (ECSD), as well as on top of GPRS, which is then referred to as EGPRS.

ECSD is capable of achieving a maximum transfer rate of 64 kbps per connection with two GSM time slots, each providing 32 kbps date rate. When compared to HSCSD, ECSD provides an identical maximum data rate but with a smaller number of GSM time slots, resulting in the conservation of transmission resources for more connections. EGPRS, on the other hand, can be viewed as a new air interface toward the GPRS system. This is due to the introduction of a higher-order modulation scheme called eight-phase-shift keying (8PSK), which yields a threefold increase in bit rate for an identical symbol rate, as opposed to Gaussian MSK (MSK) used in its predecessors. Additionally, several new techniques were first applied in EDGE, including link adaptation, hybrid automatic repeat request (HARQ) with soft combining, incremental redundancy, and advanced scheduling, before their later adoption in other 3G standards, such as Wideb and CDMA (CDMA) and Code-Division Multiple Access 2000 (CDMA2000).

EDGE utilizes convolutional coding with 1/3 rate and puncturing, along with a combination of eight phase-shift keying (8PSK) and Gaussian Minimum Shift Keying (GMSK), to support nine different levels of modulation and coding scheme (MCS) for link adaptation. When the channel condition is good, the system transmits data at a high rate, while a more robust but slower MCS option is used to improve reliability when the channel quality is poor. To reduce the overhead of radio link control (RLC) and medium-access control (MAC) headers, RLC and medium access control (MAC) protocols were merged allowing each radio frame to have a single header area for both RLC and MAC, instead of two separate headers. When an RLC/MAC frame is retransmitted, the punctured bits removed after the convolutional coder are chosen for the retransmission frame, enabling the receiver to combine redundancy information from different frames to reconstruct the original frame. The final evolution of EDGE employed more advanced techniques, including new types of modulation quadrature phase shift

keying (QPSK), 16-ary quadrature amplitude modulation (16-QAM), and even 32-ary quadrature amplitude modulation (32-QAM), in addition to 8PSK, and turbo coding.

Although EDGE reused the GSM carrier bandwidth and time slot structure, it was not restricted to GSM cellular systems. Instead, it aimed to become a generic technology facilitating an evolution of existing cellular systems toward third-generation capabilities (Furuskar et al., 1999). EDGE was designed to be compatible with North American cellular standards such as IS-54 and IS-136, which used high-order modulation to achieve data rates of up to 60 kbps by aggregating time slots. The evolution of the IS-136 standard was called IS-136HS (high speed), based on EDGE. The Universal Wireless Communications Consortium (UWCC) approved EDGE in January 1998 as the outdoor component of IS-136HS, after evaluating various proposals. The first commercial EDGE network was launched in 2003 by AT&T in the USA. With GPRS offering a maximum data rate of 115 kbps, the addition of EDGE on top of GPRS can increase this data rate to 384 kbps. This data rate is comparable with the rate for early implementations of WCDMA of the 3G cellular systems, making EDGE the bridge between the second and third generations of mobile communication systems.

3.3.4 Interim Standard 95B/IS-95B

IS-95B (Interim Standard-95B) is a second-generation CDMA cellular communication standard developed by Qualcomm as an enhancement of the IS-95 standard. IS-95B combined the standards IS-95, ANSI-J-STD-008, and TSB-74. The standardization of IS-95B was completed in 1997 under the TIA, and the world's first IS-95B commercial network was launched by a South Korean operator in 1998. The initial IS-95 system supported circuit-mode and packet-mode data services at a data rate of 14.4 kbps. Without breaking the legacy air interface design to maintain strict backward compatibility, it was upgraded to IS-95B which offered an increased data rate (Knisely et al., 1998). Through code aggregation, where a base station can assign up to eight code channels to a single mobile station, the achievable rate was increased to 115 kbps.

One of the most significant improvements was the increased capacity of the system through the introduction of the Supplemental Channel (SCH), which provided additional bandwidth to support higher data rates. The SCH was used to carry data traffic, while the Primary Traffic Channel (PCH) carried voice traffic. IS-95B also introduced the ability to use multiple PCHs simultaneously, which increased the network's capacity. It increased the number of available Walsh codes from 64 to 128, which allowed for more users to be supported on the same frequency band. Accordingly, faster and more accurate power control mechanisms helped to reduce interference and improve call quality were employed. IS-95B also improved the voice quality by introducing a new speech coding algorithm called Enhanced Variable Rate Codec (EVRC). The EVRC algorithm provided better voice quality than the previous speech coding algorithm used in IS-95. It included the ability to prioritize voice calls over data transmissions, and it supported a variety of data services, such as circuit-switched data, packet-switched data, and SMS. Another important enhancement of IS-95B was the introduction of the Message Waiting Indicator (MWI) feature, which notified users of the waiting messages on their phones.

Overall, IS-95B was a significant improvement over the original IS-95 standard, providing better voice quality, increased capacity, and better support for data services. It was widely used in North America, Asia, and other parts of the world, before being replaced by more advanced standards such as CDMA 1xRTT and EV-DO.

3.4 Key Technologies for 2G Digital Cellular

The emergence of 2G digital cellular systems represented a significant technological leap forward from the 1G analog system. These systems provided several significant advancements such as large system capacity, allowing more users to connect simultaneously, a better quality of service with fewer dropped calls and faster call setup times, better security measures that helped protect user data and privacy, more efficient use of radio frequency spectrum, and crucially, the ability to efficiently offer data services over packet-switched networks. The digitization of the cellular network allowed for the utilization of advanced techniques that were not previously possible. This section outlines the key components of the 2G digital cellular systems, including TDMA, frequency hopping, digital modulation, channel coding, speech compression, and discontinuous transmission, which empowered the success of 2G and paved the way for the development of more advanced third-generation cellular technologies.

3.4.1 Time-Division Multiple Access/TDMA

A cellular network must support a large number of active subscribers simultaneously using a finite amount of time–frequency resources. Therefore, efficient allocation of radio resources among users is a crucial design aspect for both the uplink and downlink channels since bandwidth is scarce and expensive. This allocation technique is known as *multiple access*, which involves sharing a communication channel among multiple geographically distributed users. Multiple access techniques split the signaling dimensions into frequency channels, time slots, spreading codes, etc., and then orthogonally or non-orthogonally assign them to different users.

TDMA utilizes orthogonal time slots to divide the signaling dimensions along the time axis, allowing only one user to transmit or receive in each slot. Each user occupies a repeating time slot, forming a frame that comprises several time slots. The frames consist of a preamble, an information message, and tail bits, with a repeating structure that interlaces transmissions from various users. In time-division multiple access (TDMA)/TDD systems, a portion of the time slots is reserved for the forward link, while the remaining slots are used for reverse links. In TDMA/FDD systems, identical or similar frames are used for forward or reverse transmission, but with different carrier frequencies. To eliminate the need for duplexers in the subscriber unit, TDMA/FDD systems deliberately introduce a time slot delay between the forward and reverse time slots of a specific user. The repeating frame structure in TDMA systems enables multiple users to share the same frequency band effectively.

Signal transmission for users of a TDMA system is non-continuous, resulting in lower power consumption as the transmitter can be turned off during periods when not in use most of the time. Non-continuous transmission simplifies the system design where channel estimation can be performed during the time slots of other users, and the handover process is much simpler for a mobile phone that is able to listen to other base stations during idle time slots. Another advantage is that a TDMA system can flexibly assign time slots, facilitating the implementation of high data rates by aggregating multiple slots for a single user, and supporting unbalanced traffic by dynamically assigning slots between downlink and uplink. One major challenge in a TDMA system is the synchronization of the uplink channel. The signals transmitted from spatially separated users induce different propagation delays. Moreover, mobile users may move continuously, and the multi-path propagation environment also varies, making the synchronization of the uplink channel hard to achieve. The bandwidth of a TDMA channel is generally larger than that of an FDMA channel in the previous analog system. If the signal bandwidth exceeds the coherent bandwidth of the wireless channel, inter-symbol interference (ISI) raises, and therefore an equalizer is required at the receiver to compensate for the ISI.

Using the GSM standard as an example, it employed TDMA combined with FDMA as the multiple access scheme. GSM was allocated paired frequency bands with a bandwidth of 25 MHz each: 890–915 MHz for the uplink transmission and 935–960 MHz for the downlink transmission. To avoid interference with other systems operating in the neighboring frequency bands, a pair of guard bands was applied, and the remaining spectrum was divided into a total of 124 FDMA narrowband channels, each with a bandwidth of 200 kHz. A TDMA channel multiplexes eight time slots with a duration of approximately 0.577 ms per slot, and the content carrying in a time slot is called a burst. To compensate for synchronization error and multi-path delay spread, a guard period is inserted at the tail of each burst. Several burst types were defined for different functions, including normal burst, frequency correction burst, synchronization burst, access burst, and dummy burst. Eight bursts with a total length of 4.615 ms form a TDMA frame, which cyclically repeats.

3.4.2 Frequency Hopping

Frequency hopping has several benefits. First, it can improve the quality of the signal by reducing co-channel interference from other wireless devices operating on the same frequency band. Second, it can enhance the security of wireless communication by changing frequencies in a predefined pattern, where the transmitted signal becomes more difficult for unauthorized parties to intercept and detect the transmitted signal. Finally, frequency hopping can effectively mitigate the effect of multi-path channel fading by transmitting signals over a bandwidth much larger than the width of the original signal, resulting in a gain of wideband signal transmission similar to that of spectrum spreading. This technology was invented by the film star Hedy Lamarr and the composer George Antheil during World War II and released in their patent *Secret communication system* (Markey & Antheil, 1941). Frequency hopping works by continuously changing the carrier frequency

of the transmission from one time slot to another, with the hopping sequence determined by a mathematical algorithm shared between the transmitter and the receiver. A frequency synthesizer at the transmitter generates the hopping frequency carriers according to a pseudorandom sequence termed the spreading code, and the same spreading code is used at the receiver to generate the frequency carriers to down-convert the received signals. There are two types of frequency hopping. If the hop time exceeds a symbol period, it is called slow frequency hopping. Otherwise, it is the fast frequency hopping when the hop time is less than a symbol period. GSM adopted slow frequency hopping, taking advantage of the inherent frequency agility of the transceivers that can transmit and receive on different channels. The carrier frequency changes every TDMA frame at a prescribed rate of 217 times per second.

3.4.3 Digital Modulation

Digital modulation offers several advantages over analog modulation adopted in 1G analog cellular systems. First, it provides high spectral efficiency, which allows more information to be transmitted over a given bandwidth than analog modulation. Second, digital components are typically more power-efficient than analog devices, meaning it requires less power to transmit the same amount of information. Third, digital modulation can be combined with other digital processing techniques such as channel coding, equalization, and spread spectrum to effectively and efficiently resist hardware impairments, multi-path fading, additive noise, and interference, resulting in enhanced robustness against channel impairments. Fourth, the information bits carried on modulation constellations are much easier to encrypt than analog signals, resulting in a high level of security and privacy. As a digital system, the GSM standard has taken advantage of digital modulation, where GMSK was adopted to modulate the information bits. It was determined over other modulation schemes as a compromise among spectral efficiency, transmitter complexity, and limited spurious emissions. One particular advantage of GMSK is a constant-modulus signal (constant envelope signal), which alleviates the nonlinear distortion problem caused by a power amplifier. As an enhancement of GSM, EDGE employed higher-order phase-shift keying and M-ary QAM to further improve the data transmission rate.

3.4.4 Channel Coding

Channel coding is a critical technique used in mobile communications to improve the reliability of data transmission over a noisy channel. In wireless communication systems, the signal can be distorted by various factors such as interference, fading, and noise, which can cause errors in the received data. Channel coding helps to protect the transmitted data against such errors by adding redundant information to the original data before transmission. This redundant information allows the receiver to detect and correct errors in the received data. The process of adding redundant information is called encoding, and the process of detecting and correcting errors is called decoding. Wireless communication systems generally need various channel coding schemes for forward error correction such as Reed–Solomon codes, convolutional codes, and turbo codes, along with error detection techniques such as the extensively used cyclic redundancy check (CRC) codes.

The 1G system suffered from worse voice quality since the induced noise and interference on the analog transmitted signal cannot be filtered out. The 2G digital system enabled the adoption of channel coding to improve performance (AMPS applied channel coding but only for the control channel). Contrary to speech coding which tries to compress the amount of data as little as possible, channel coding intentionally adds redundancy bits to the original information to detect or correct errors incurred during the transmission. In GSM, for example, different levels of protection are provided for the speech bits, according to their importance. The speech codec of a GSM terminal generates a block of 260 bits every 20 ms. From the users' perspective, the perceived speech quality depends more on some part of this block rather than every bit equally. Consequently, the block is divided into three parts: Class Ia: 50 bits—most important, Class Ib: 132 bits—moderately important, and Class II: 78 bits—least important. First, CRC is added at the tail of Class Ia for error detection. These 53 bits, together with Class Ib and a 4-bit tail sequence, amounting to a total of 189 bits, are input into a convolutional encoder with a rate of 1/2 and a constraint length of 4. The remaining bits in Class II are added into the encoded sequence without any protection. Finally, every 20 ms speech signal is transformed into a transmission block with a length of 456 bits, resulting in a coding rate of 22.8 kbps.

Moreover, interleaving is used in conjunction with channel coding to improve the reliability of data transmission. Interleaving is applied to mitigate the effects of burst errors caused by fading and interference in wireless channels. Burst errors are characterized by multiple consecutive bit errors that occur within a short period of time, which can cause significant

data loss and affect the quality of the communication. Interleaving is achieved by dividing the data into small blocks and transmitting them in a different order than the original sequence. The channel coding is applied to each block of data before interleaving, and the interleaved blocks are transmitted over the wireless channel. The receiver then applies the channel decoding to the received interleaved blocks, de-interleaves them, and reconstructs the original data sequence.

3.4.5 Speech Compression

The most basic service provided in early mobile systems was voice transmission. In the 1G cellular system, the speech signal was modulated directly over carrier frequency using analog modulation. The digitization of 2G requires digitizing analog speech signals before transmission. A classical speech processing technique known as Pulse Coded Modulation (PCM) was employed in wireline telephone systems to code speech signals for transmission over high speed backbone or optical fiber lines, but its coding rate of 64 kbps was too high for air interface transmission due to radio resource constraints. To solve this problem, the GSM working group considered various speech coding and synthesis algorithms to reduce redundancy in the sounds of the voice. They evaluated these algorithms based on subjective speech quality, processing delay, power consumption, and complexity. After careful consideration, they selected the RPE-LTP (which means Regular Pulse Excitation Long-Term Prediction) codec. The RPE-LTP codec works by using a number of past samples to predict the current sample, resulting in a lower coding rate. The speech signal is sampled every 20 ms to obtain 260-bit data blocks, equivalent to a coding rate of 13 kbps, making it the preferred choice for the GSM system.

3.4.6 Discontinuous Transmission/DTX

The key idea of discontinuous transmission is suspending signal transmission during the interval called the silence period, based on the observation that individuals typically speak for less than half of the time during a conversation over a mobile phone. DTX can significantly decrease power consumption in mobile devices, as the transmitter is turned off during silent intervals, thereby extending overall battery life and allowing for longer conversations. In addition to reducing energy usage, it can also enhance service quality by mitigating mutual interference among mobile devices operating on the same frequency channel. As DTX limits the duration of transmitter activity, it enlarges the capacity of a mobile network by carrying the singles of more users over the same frequency channel. This technique was first introduced into TDMA-based digital cellular systems since a 1G analog user occupies its assigned FDMA channel for the whole time.

To implement the silence suppression function in mobile communication systems, two primary components are necessary: voice activity detection (VAD) and a comfort noise generator. VAD is used to distinguish between periods of speech activity and silence in the audio stream transmitted from the mobile device. It analyzes the signal's energy level and spectral characteristics to determine whether the signal contains speech. During periods of speech, the VAD identifies the presence of vocal activity and transmits the corresponding speech data to the base station. However, during periods of silence, the VAD detects the absence of speech and suspends signal transmission to prevent the base station from wasting transmission resources. In case a voice signal is mistakenly identified as noise, the transmitter may turn off, resulting in an unpleasant effect known as clipping. Unlike analog voice signals, a digital system suffers from absolute silence during the turn-off period of the transmitter. The users' subjective perception might be very annoying on the reception side because it seems that the connection drops. To address this issue, the receiver generates a small comfort signal that mimics the background noise during the silence period, improving the user's subjective perception of the connection quality.

3.5 Summary

This chapter delved into the key driving forces that led to the transition from 1G to 2G and introduced a variety of standards that constitute 2G. Key technologies that propelled the evolution of 2G were discussed, in an effort to provide readers with a comprehensive understanding of this crucial phase in the evolution of cellular communications. Through this comprehensive overview, the readers gain an overview of this generation. To enable an insight into the evolution and most importantly gain a deep understanding of the operation of a cellular network, we will select the most dominating 2G standard, i.e., GSM, as an example to offer an in-depth exploration of a 2G mobile communication standard.

3.6 Exercises

1. Can you identify some drawbacks of the first-generation cellular technology? What causes these drawbacks?
2. What is the most technological difference between 1G and 2G? Describe the advantages of the 2G technology.
3. What were the initial frequency bands assigned to the GSM deployment? What are the differences between GSM-900 and DCS-1800?
4. Select the multiple access technologies adopted by 2G digital cellular standards:
   (A) Frequency-division multiple access (FDMA)
   (B) Time-division multiple access (TDMA)
   (C) Code-division multiple access (CDMA)
   (D) Orthogonal frequency-division multiple access (OFDMA)
5. After the commercial deployment of GSM, the mobile industry invested a lot to enhance digital cellular systems. This resulted in widely accepted technologies like GPRS and EDGE. Describe the driver of developing 2.5G cellular standards.
6. What is the main difference between circuit-switched and packet-switched networks?
7. Why do wireless communications need channel coding? Identify two basic forms of channel coding.
8. The speech codec of a GSM terminal generates a block of 260 bits every 20 ms. The block is divided into three parts: Class Ia: 50 bits—most important, Class Ib: 132 bits—moderately important, and Class II: 78 bits—least important. First, CRC is added at the tail of Class Ia for error detection. These 53 bits, together with Class Ib and a 4-bit tail sequence, amount to a total of 189 bits. The bit sequence is input into a convolutional encoder with a rate of 1/2, converted into 378 bits. The remaining bits in Class II are added into the encoded sequence without any protection. Calculate the coding rate.
9. You observe *the silence period* that individuals typically speak for less than half of the time during a voice conversation over a mobile phone. Design a technique to take advantage of the silence period to improve the performance of a cellular communication system. What are the benefits obtained?

4 Global System for Mobile Communications (GSM)

4.1 Frequency Bands and Key Features

Initiated in 1982, the GSM committee was assigned to the task of specifying a Europe-wide cellular system that overcomes the limitations of the legacy 1G systems. The initial conception by the CEPT was dedicating the new system to the 900 MHz band. Later on, during the further development and international deployment, upon requests by different countries and regions that had joined onboard, GSM was adopted to operating at more frequency bands to be tailored to the various requirements of local networks, as listed in Table 4.1.

As the first digital cellular standard and the first internationally deployed cellular system, GSM outperforms its 1G analog predecessors in various aspects. First, digital transmission is much more robust than analog technologies against frequency-selective distortion and additive noise, leading to a significant enhancement in voice quality. Second, digital transmission technologies outperform analog ones in spectral and power efficiency, which implies increased network capacity, faster data transmission, and improved battery life. Third, advanced security mechanisms such as encryption and SIM are enabled, which provide better protection against eavesdropping and fraud. Fourth, some basic data services and new features are introduced, such as SMS, caller identification (ID), call waiting, call forwarding, and conference calling, which were not supported by the analog systems. Last but not least, thanks to the deployment of SIM, IMEI, the global standardization of network architecture and protocols, and the introduction of roaming clearinghouses, the establishment of international roaming agreements became possible, allowing users for the first time to directly use their phones while traveling abroad.

4.2 GSM Architecture

4.2.1 System Architecture

Generally, a GSM system generally consists of two components: the mobile stations (MSs) and the fixed infrastructure that is commonly known as the Public Land Mobile Network (PLMN). More specifically, the PLMN can be divided into three subsystems from the architectural perspective, as illustrated in Fig. 4.1: (i) the BSS, which is the GSM radio network that establishes efficient, reliable, and secure radio links between the MSs and the fixed infrastructure, (ii) the NSS, which is the GSM backbone network that manages the call routing and switching between different MSs and networks, and (iii) the OSS, which manages the operation and maintenance of the network, including monitoring, testing, and configuration management.

4.2.1.1 Mobile Station (MS)

MSs are the devices used by mobile service subscribers to access the services. Each MS consists of two modules: the SIM and the mobile equipment (ME). The SIM stores subscriber-specific information, which includes the identity, authentication keys, the phone number, SMS messages, and contacts. It is mostly implemented in form of a removable and exchangeable smart card, but sometimes also as a fixed installed chip (known as the plug-in SIM). The ME, in contrast, provides all functionalities that are independent from the subscriber identity. More specifically, it can be further divided into two major components: the terminal equipment (TE) that performs application-specific functions that are independent from the GSM system (e.g., a pager) and the mobile terminal (MT) that executes functions related to the GSM radio interface. Additionally,

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5_4

Table 4.1 Different GSM bands and their regional deployments

GSM band	Frequency (MHz)	UL (MHz)	DL (MHz)	Regional deployments
GSM-850	850	824.2–848.8	869.2–893.8	Canada, the USA, Caribbean, Latin America
GSM-900	900	890.0–915.0	935.0–960.0	Europe, Africa, Asia, Oceania
DCS-1800	1800	1710.2–1784.8	1805.2–1879.8	Europe, Asia, Africa, Oceania
PCS-1900	1900	1850.2–1909.8	1930.2–1989.8	Canada, the USA, Caribbean, Latin America, some parts of Asia and Oceania

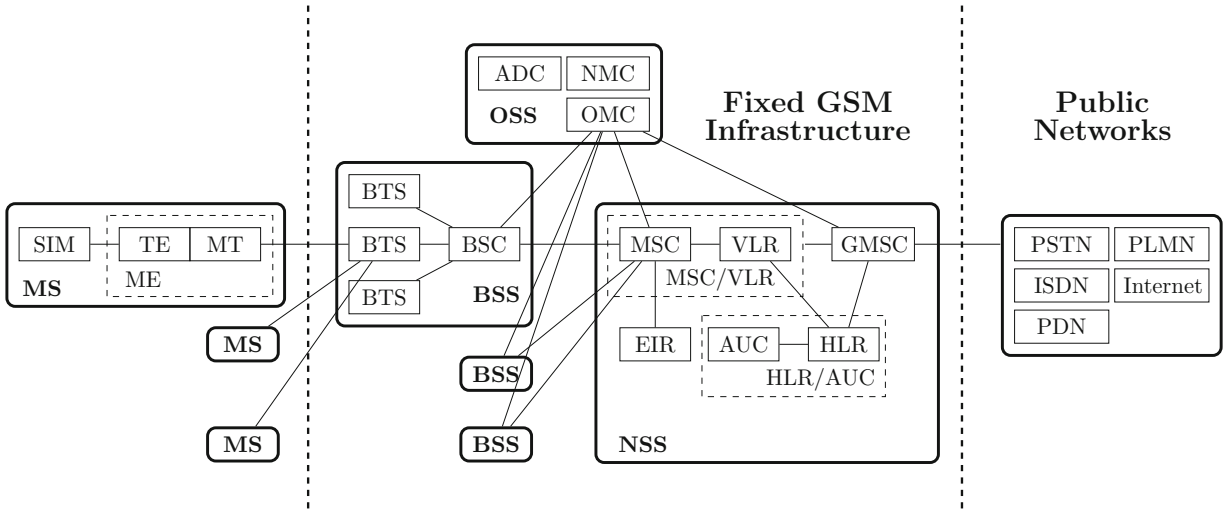


Fig. 4.1 The overall GSM PLMN architecture

there can be an optional component called the terminal adapter (TA), which acts as a gateway between the TE and the MT to provide necessary interface for ensuring compatibility.

Since GSM services generally rely on identification and authentication procedures, an ME cannot obtain the network usage privileges to become functioning unless combined with a SIM. This separation of SIM and ME was intentionally standardized by GSM as a key feature, for the purposes of (i) decoupling the subscriber mobility from the equipment mobility and (ii) allowing the subscriber to keep a consistent personalization of their service and data independently from the mobile terminal.

4.2.1.2 Base Station Subsystem (BSS)

In GSM, every MS is connected with the fixed infrastructure at a BSS—more specifically, it is connected via the radio interface U_m to a BTS, which is concerned with the radio transmission and reception functions in the cell area around it. While assigned with all the radio frequency (RF) and baseband signal processing components for user data traffic, BTS in GSM has little to do with controlling except for only few protocol functions suchlike the Link Access Protocol on the Dm Channel (LAPDm). The essential functions of control and management for radio resource management (RRM) are resided in the BSC, which is connected to multiple BTS over the A-sub interface, controlling them and coordinating complex cross-BTS tasks, suchlike radio channel allocation, handover procedure, and transmission power control. Lying on a higher hierarchical level above the BTS, every BSC assembles with its subordinate BTS—typically up to 40 of them (Steele et al., 2001)—into a BSS. This hierarchical design of BSS with the functions split between BTS and BSC is intended for keeping the base station (BS) small and cheap.

4.2.1.3 Network Switching Subsystem (NSS)

Each BSS is connected over the A-bis interface (a.k.a. the A interface) to an MSC, which is the key node of the NSS concerned with all necessary switching functions of GSM, such as routing path search, signal routing, and service feature processing. A typical MSC controls dozens of BSCs and has a capacity of several tens of thousands of subscribers (Steele et al., 2001). Different MSCs are connected with each other over the E interface, constructing the main body of the GSM

backbone network. Usually, one or several MSCs are selected as gateway MSC (GMSC) to provide the interface between the PLMN and external networks.

Generally, MSCs play in GSM PLMNs a similar role like switching exchanges do in fixed networks. However, compared to fixed networks, GSM has to include some additional functions to support and manager user mobility. Several databases are therefore introduced into the GSM NSS, namely the HLR, the VLR, the EIR, and the AuC.

Both HLR and VLR are defined to maintain subscriber information. The HLR is a PLMN-specific centralized database: Typically, there is only one HLR in each GSM PLMN, which is connected to all MSCs via the C interface. It stores the two types of information of every subscriber in the network: the subscription information and the current location. However, a subscriber may move to another country/region other than that of its home network: In this case, the subscriber may have to visit a foreign GSM PLMN, in which its information is not contained in the HLR. To recognize and authenticate the visiting subscriber so that GSM services can be provided and calls can be handled, this visited network will have to access the HLR of the visiting subscriber's home network. A frequent and continual access of this kind would generate a significant inter-GSM traffic that is expensive. To tackle this issue caused by subscriber mobility, VLR is designed in to provide a local copy of such visiting subscribers. A VLR is associated with one or multiple MSCs, which jointly cover a so-called VLR area. When a subscriber visits a VLR area, the network copies its subscription information from the subscriber's HLR to the associated VLR. Under each VLR area are a number of location areas, each consisting of one or multiple cells or sectors. When a subscriber enters a location area, it triggers the location update (LU) procedure so that the network updates its current position at the associated VLR. Thus, when an incoming call arrives, the called subscriber will be paged in all cells/sectors associated with location area of its latest LU to efficiently find it. When the subscriber moves into a new location area that is dedicated to a VLR area other than the old one, its subscription information will also be copied from the HLR to the new VLR. A VLR is connected to the HLR via the D interface, to its controlled MSCs via the B interface, and with other VLRs via the G interface.

The other two databases, i.e., AuC and EIR, are responsible for the GSM security in different aspects. Basically, the AuC is closely associated with the HLR to generate and store confidential data and keys, including the subscriber's secret K_i key and the $A3$ and $A8$ security algorithms, which serve for user authentication and radio interface encryption. The only entity that AuC communicates with is the HLR, and the interface between them is called the H interface. The EIR, on the other hand, stores the IMEI of the MEs to allow the tracking of stolen and malfunctioning terminals. Three IMEI lists are stored there, namely the white list of the MEs approved in the GSM network, the black list of those banned from the network, and the gray list of those that must be tracked. The EIR communicates only with the MSC, through the F interface.

4.2.1.4 Operation and Support Subsystem (OSS)

To ensure efficient management, monitoring, maintenance, and control to the network, several additional functional blocks are defined in the GSM system, including the OMC, the network management center (NMC), and the administration center (ADC), which assemble into the OSS. Especially, each OMC is typically dedicated to a BSS or an NSS. In some deployments, the AuC and the EIR are considered not as part of the NSS, but together with the OMC as components of the Operation and Maintenance Sybsystem (OMSS) (Garg, 2010; Eberspächer et al., 2001).

4.2.2 Identifiers and Addressing

In GSM systems, the identity and location information of the subscribers and the mobile equipments are managed with a number of identifiers, as listed below (Fig. 4.2).

4.2.2.1 IMEI

To internationally identify MSs, a unique IMEI is issued by the equipment manufacturer to every individual MS and registered at a GSM network in its EIR for management, as explained in Sect. 4.2.1.3. Each IMEI consists of four parts, namely (i) the centrally assigned 6-decimal type approval code (TAC), (ii) the manufacturer-assigned 6-decimal final assembly code (FAC), (iii) the manufacturer-assigned 6-decimal serial number (SNR) , and (iv) the 1-decimal spare (SP).

4.2.2.2 IMSI

Similar to but separated from the MS, every subscriber is also internationally uniquely identified with an IMSI, which is stored in the SIM. Each IMSI consists of three parts, namely (i) the internationally standardized 3-decimal mobile

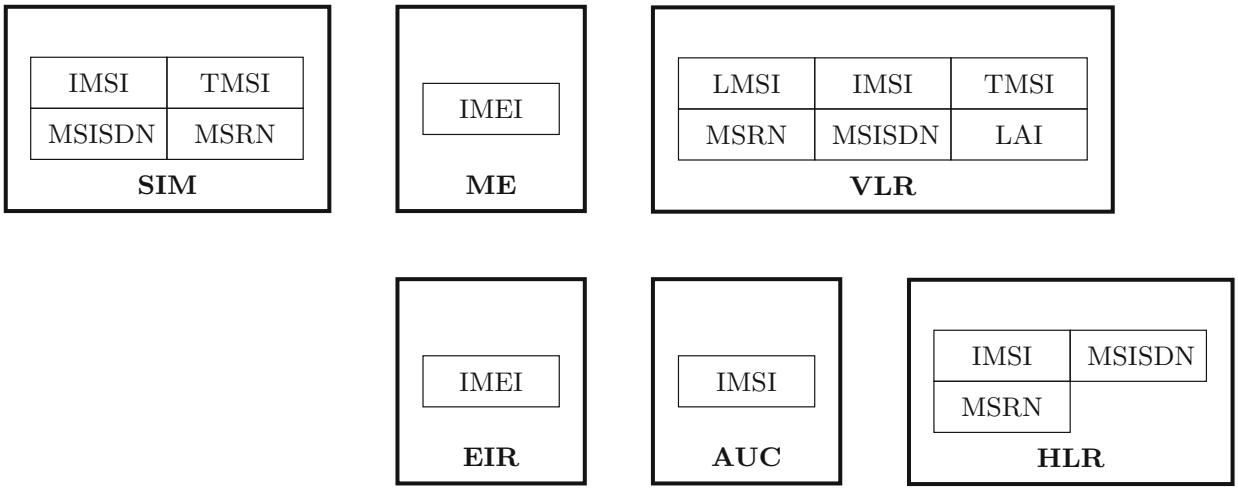


Fig. 4.2 The storage of identifiers in GSM

country code (MCC), (ii) the country-certified 2-decimal mobile network code (MNC), and (iii) the network-assigned mobile subscriber identification number (MSIN) with up to ten decimal places.

Thus, by connecting a SIM to an ME, the IMSI is correspondingly coupled with the IMEI, and the GSM network is only able to correctly serve and bill the subscriber when both the IMSI and the IMEI are valid.

4.2.2.3 MSISDN

One novel feature that was introduced by GSM for the first time is the separation of subscriber identity, which is identified by means of IMSI, from its "call number," which is the mobile subscriber ISDN number (MSISDN). The reason of this separation is twofolded. First, since the callers only need the mobile subscriber ISDN number (MSISDN) of a specific subscriber to call them, it protects the IMSI that contains the confidential identity information from exposure to the public. Second, each subscriber can be assigned to multiple MSISDN by the network operator, which can be used to separate different GSM-supported services for the same subscriber from each other, e.g., voice, data, and fax.

An MSISDN consists of three parts: (i) the internationally standardized country code (CC) of up to three decimal places, (ii) the national destination code (NDC) of typically 2–3 decimal places, and (iii) the subscriber number (SN) with up to ten decimal places. It is centrally stored in the HLR.

4.2.2.4 MSRN

As explained in Sect. 4.2.1.3, to enable international roaming and to ease the mobility management, the subscription information of a subscriber shall be copied from the HLR of to the local VLR. In this process, the VLR assigns the subscriber a temporary location-dependent ISDN number that is called the mobile station roaming number (MSRN) and has exactly the same structure of CC + NDC + SN as the MSISDN does. The assignment of VLR can be triggered both by the LU process when a subscriber enters a new location area (LA) and upon request by the HLR for setting up connection for an incoming call.

4.2.2.5 TMSI and LMSI

As a temporary MSRN is assigned regarding the MSISDN, the IMSI also has its temporary replacement that is only locally valid, which is the temporary mobile subscriber identity (TMSI). The TMSI is assigned to the subscriber by the local VLR and stored by the MS in the SIM. It remains valid only during the period of the MS visiting the area of the VLR and can even be changed during that period. By using the TMSI instead of the IMSI to identify and address the MS, the identity of the subscriber is completely hidden from the radio interface, which protects the privacy and enhances the GSM security.

Being only locally valid, the TMSI is not internationally unique. However, this deficiency can be made up for by combining the TMSI with the LA identity (LAI) that identifies the LA internationally uniquely (see Sect. 4.2.2.6).

Furthermore, a local identifier called local mobile subscriber identity (LMSI) can be assigned by the VLR to each MS within its area when the latter registers with the VLR. Once assigned, the LMSI is also sent to the HLR. It is not used by

the HLR itself, but appended to every message from HLR to the VLR, which is then used as a searching key to enhance the database operation efficiency.

4.2.2.6 LAI, CI, SPC, and BSIC

In GSM, not only the terminals and the subscribers but also the areas and network nodes are identified.

First of all, each LA of a PLMN is assigned with a LAI, which consists of the CC, the MNC, and an LA code (LAC) with up to five decimals.

Within each LA, every cell is identified by cell identifier (CI) with up to 2×8 bits. Appending it to the LAI of the LA that it is associated with, it constructs the global cell identity (GCI) that internationally uniquely identifies every cell.

In GSM PLMN, MSCs and location registers are addressed with ISDN numbers. Additionally, to address them in the signaling network, each of them can also be assigned with a signaling point code (SPC).

At last, for radio resource management, it is essential to distinguish neighboring base stations which are close to each other. For this purpose, each BTS has a Base Transceiver Station Identity Code (BSIC) which consists of two parts: a 3-bit network color code (NCC) and a 3-bit base transceiver station color code (BCC). Directly adjacent PLMNs must have different NCC, and directly adjacent BTSs must have different BCC.

4.3 Radio Interface

4.3.1 Logical Channels

Logical channels in GSM can be generally divided into (TCHs) and signaling channels. More specifically, there are two kinds of traffic channels: the mobile B (Bm) channels that transmit in full rate and the lower-rate mobile (Lm) channels that transmit in half rate. All TCHs are bidirectional. Depending on the functionality, different signaling channels can be UL-only, DL-only, or bidirectional, as listed in Table 4.2.

4.3.2 Physical Channels

4.3.2.1 Modulation

GSM uses GMSK for its air interface, which is a modified version of minimum-shift keying (MSK). MSK is a special continuous-phase FSK that keeps the quadrature component delayed by half the symbol period w.r.t. the in-phase component and uses a half sinusoidal impulse to encode each bit. Compared to other digital modulation schemes such as QPSK and offset QPSK (OQPSK), MSK has better spectral efficiency and less out-of-band distortion. This spectral efficiency is even further improved in GMSK, which uses a Gaussian-impulse filter to pre-shape the digital impulses before applying the frequency modulator. As a cost, the Gaussian filter also introduces ISI, which is resolved in GSM by a *Viterbi* channel equalizer at the receiver.

Table 4.2 GSM logical signaling channels

Group	Channel	Direction
Broadcast Channel (BCH)	Broadcast Control Channel (BCCH)	DL
	Frequency Correction Channel (FCCH)	
	Synchronization Channel (SCH)	
Common Control Channel (CCCH)	Random Access Channel (RACH)	UL
	Access Grant Channel (AGCH)	DL
	Paging Channel (PCH)	DL
	Notification Channel (NCH)	DL
Dedicated Control Channel (DCCH)	Stand-alone Dedicated Control Channel (SDCCH)	UL + DL
	Slow Associated Control Channel (SACCH)	
	Fast Associated Control Channel (FACCH)	

4.3.2.2 Radio Frame, Multiple Access, Duplex, and Frequency Hopping

GSM uses a hybrid multiple access scheme that combines FDMA and TDMA and a hybrid duplex mode that combines FDD and TDD. It reserves two 25 MHz frequency bands: taking the GSM-900 band as example, which is 890 MHz to 915 MHz for UL and 935 MHz to 960 MHz for DL, respectively. Each of the two bands is divided into 124 channels of 200 kHz bandwidth and two 100 kHz guardbands. Each of these 200 kHz channels is organized in radio frames, where each frame is divided into eight time slots and every slot carries a TDMA channel. The channels in UL and DL are paired and synchronized to construct an FDD bidirectional link, while for each pair the UL slot is always delayed by three slots w.r.t. its corresponding DL slot, which implements a TDD for every paired link, as illustrated in Fig. 4.3. The motivation behind such design is to maximize the spectral efficiency at the BTS while reducing the implementation cost of MSs by getting rid of the expensive duplex front-end components.

Each TDMA frame lasts 1250 bit periods, which is about 4.615 ms, i.e., every slot is about 576.9 µs long. When used, a slot transmits a so-called data *burst*, which is the basic information unit transmitted over GSM air interface. Depending on the channel type, there are in total five types of bursts with different data formats, namely (i) the normal burst used for TCH and control channels other than RACH, (ii) the frequency correction burst used for FCCH, (iii) the synchronization burst used for SCH, (iv) the dummy burst used for BCCH, and (v) the access burst used for random access to the RACH without reservation. Each burst type has its specific data structure, as illustrated in Fig. 4.4 (3GPP TS 45.002: GSM/EDGE multiplexing and multiple access on the radio path, V17.0.0, 2022).

Furthermore, to overcome the frequency-selective channel fading caused by multi-path propagation, GSM provides an optional frequency hopping (FH) that allows a TCH to change its carrier frequency with each burst, i.e., 217 hops per second. This FH enhances the effective signal-to-noise ratio (SNR) to achieve a better voice quality.

Specifically, GSM uses a hopping sequence of 1, 3, 4, or 8 frequencies, which are selected from a pool of 64 available frequencies. The FH is carried out in such a synchronized manner that neither the frequency separation nor time delay between the UL and DL channels for the same UE is impacted (see Fig. 4.5). The use of FH is upon decision of the network, and the hopping sequence is generated based on the identities of the subscriber and the BTS.

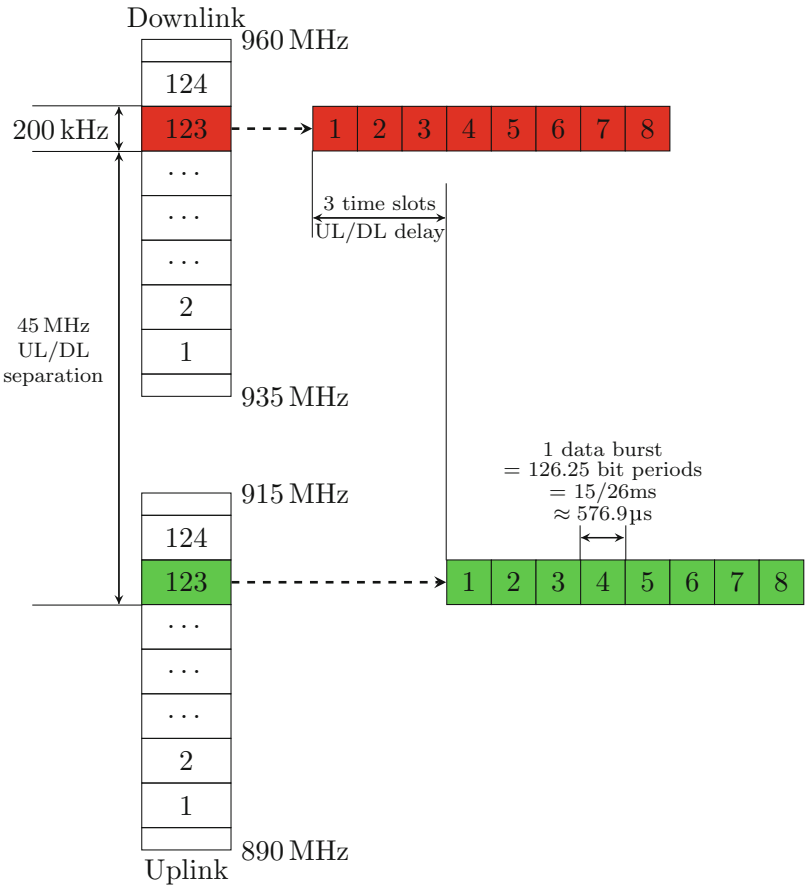


Fig. 4.3 A GSM frame in the GSM-900 band

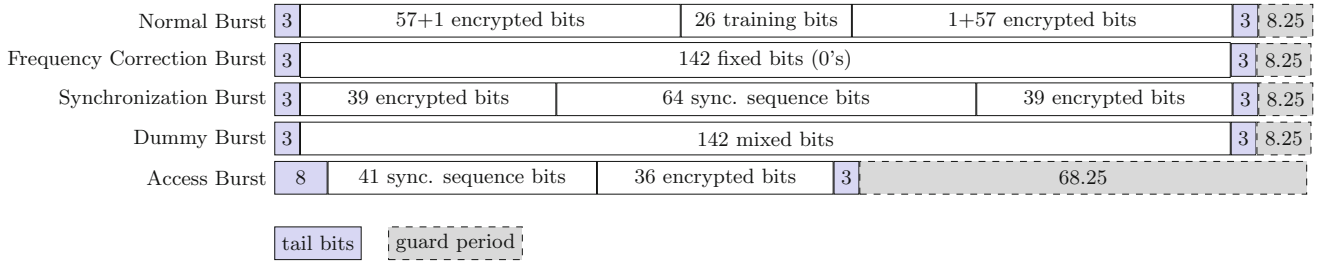


Fig. 4.4 Bursts of the GSM TDMA procedure

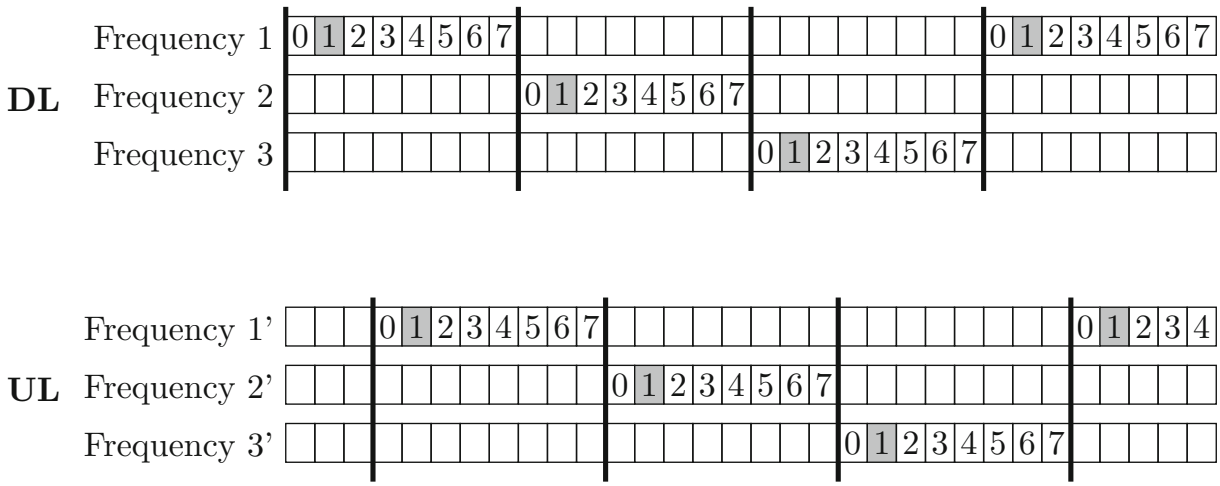


Fig. 4.5 An example of FH in GSM with three hopping frequencies

4.3.2.3 Coding and Error Protection

Mainly specified for speech service, GSM has a set of carefully designed mechanisms for speech coding and error protection.

Generally, there are three different kinds of speech codecs, namely (i) waveform codecs, which treat the speech signal as a vibration waveform and aim at reproducing the waveform at the receiver as good as possible, (ii) vocoders, which are parametric digitizers specified for human speech and leverage certain properties of the human speech production, and (iii) hybrid codecs, which combine both the aforementioned techniques. While waveform codecs outperform vocoders with an excellent speech quality under high data rates, they fail to compromise with low data rates. On the other hand, vocoders have a low upper bound of voice quality even under high data rates. Hybrid codecs, as a combination solution, have an intermediate requirement for minimal data rate and deliver fair speech quality. Due to the limited data rate and cost concerns, GSM deploys vocoder as its speech coding solution. More specifically, it uses Regular Pulse Excitation - Long Term Prediction (RPE-LTP), a specific linear predictive coder (LPC) which models the speech signal as a set of linear predictive filters that represent the spectral envelope of the speech signal, and uses a combination of fixed and adaptive codebooks to represent the filter coefficients.

Speech signals in GSM are first sampled at a rate of 8000 samples per second and linearly quantized with 13 bits per sample. This forms a 104 kbps data stream, which is then taken by the RPE-LTP coder and compressed to a rate of 13 kbps, generating a data block of 260 bits in every 20 ms. This significant compression allows for more efficient use of the limited radio spectrum available to GSM networks.

To protect the transmitted bursts from errors, a multi-stage channel coding and interleaving scheme is designed for GSM in ETSI TC-SMG (1996). Taking the full rate TCH (TCH) as example, each 260-bit speech block generated by the speech coder is divided into two parts: 182 most important bits and 78 unimportant bits. Out of the 182 important bits, 50 are further classified as the most important ones, which are block coded to generate 3 extra bits of cyclic redundancy and appended with the rest 132 important bits and 4 tail bits, totaling to 189 bits. These 189 bits are further encoded by a convolutional coder with constraint length $k = 5$ and coding rate of $1/2$, converted into 378 bits. Finally, the 78 unimportant bits are appended to generate a 456-bit output of the channel encoder. These bits are then reordered and partitioned into eight blocks

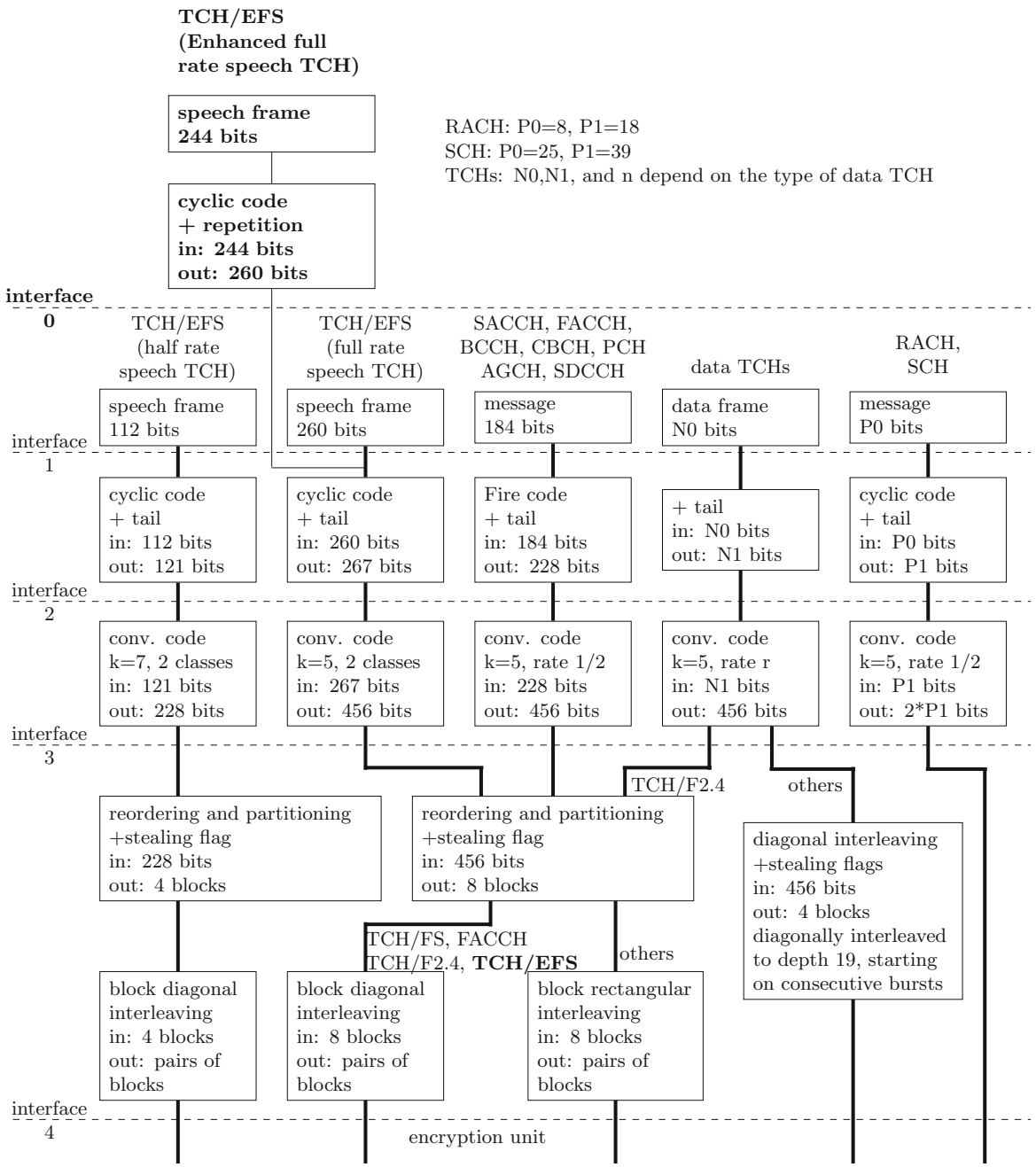


Fig. 4.6 Channel coding and interleaving organization of GSM (ETSI TC-SMG, 1996)

with stealing flags added, before undergoing a block diagonal interleaving. In the end, pairs of blocks are generated and sent to the encryption unit. Similar procedures are valid for other logical channels, with a slight variation in the specifications (Fig. 4.6).

4.4 Security

To protect both subscribers and MNOs from fraudulent activities, multiple security measures are applied in GSM. Generally, three different security aspects are involved: identification, authentication, and ciphering.

4.4.1 Identification Protection

As described in Sect. 4.2.1.1, important subscriber information is stored in the SIM. To protect such information in case the SIM is stolen, a personal identification number (PIN) of 4–8 decimal digits is assigned to every SIM. After power-up of the ME or re-inserting the SIM, the user will be required to enter the PIN to activate the SIM. After three consecutive attempts with incorrect PIN, the SIM will be blocked and can only be unblocked by a preset 8-digit PIN unblocking key (PUK), which is also stored in the SIM. After ten consecutive attempts with incorrect PUK, the SIM will be permanently blocked.

On the network side, as introduced in Sect. 4.2.2, the user identity is protected by using the TMSI instead of IMSI to achieve a pseudo-anonymity over the air interface, which is openly exposed to potential eavesdroppers.

4.4.2 Authentication

The authentication of a subscriber in GSM is realized through a mutual key verification mechanism, which is essentially based on a symmetric encryption algorithm known as A3. The A3 algorithm takes two 128-bit numbers as input to generate a 32-bit output code. In GSM, every subscriber is assigned with a 128-bit Subscriber Authentication Key (Ki), when it is added to a home network for the first time. This key is stored in both the SIM and the AuC, associated with the subscriber's IMSI. Each time when an authentication is needed, the AuC challenges the subscriber by sending it a 128-bit Random Number (RAND). The A3 algorithm is then executed at both the MS and the AuC, taking the Ki and the RAND to generate a 32-bit Signature Response (SRES). The MS-generated SRES (based on the SIM-stored Ki) is then sent back to the VLR and compared with the AuC-generated SRES (based on the AuC-stored Ki) there. Only when both SRES are identical, the authentication is successfully passed. The principle is briefly illustrated in Fig. 4.7.

4.4.3 Ciphering

In concerns of security and privacy, in GSM systems, both signaling and payload data are transmitted ciphered over the air interface. As a fundamental of the ciphering and deciphering processes, a Ciphering Key (Kc) is generated on both MS and

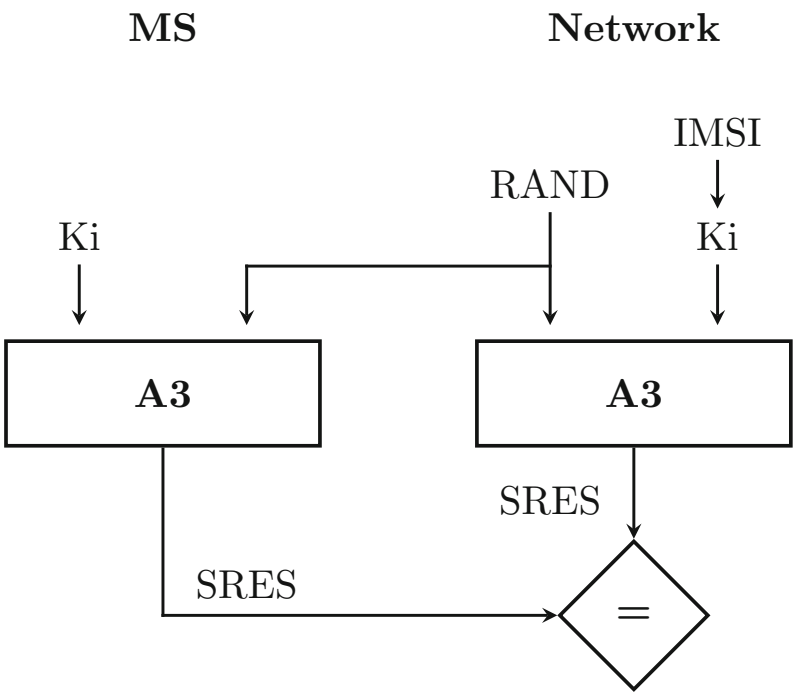


Fig. 4.7 GSM subscriber authentication

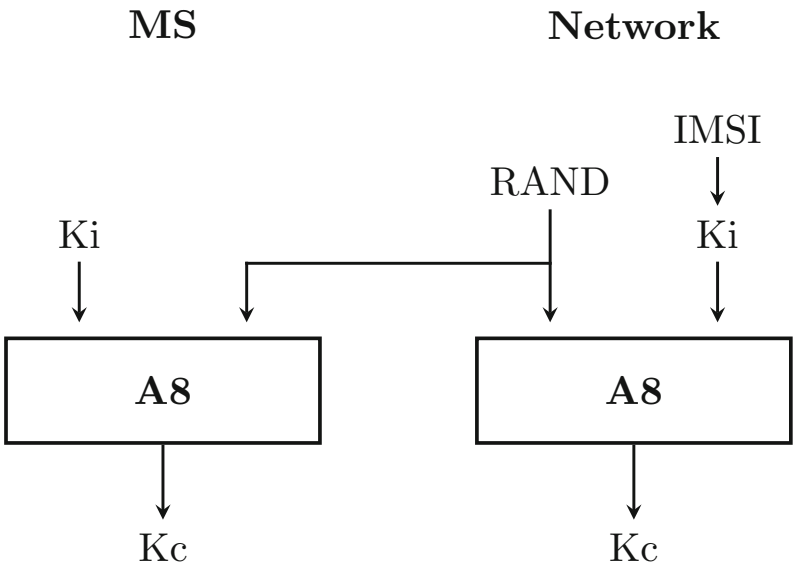


Fig. 4.8 Generation of the GSM cipher key Kc

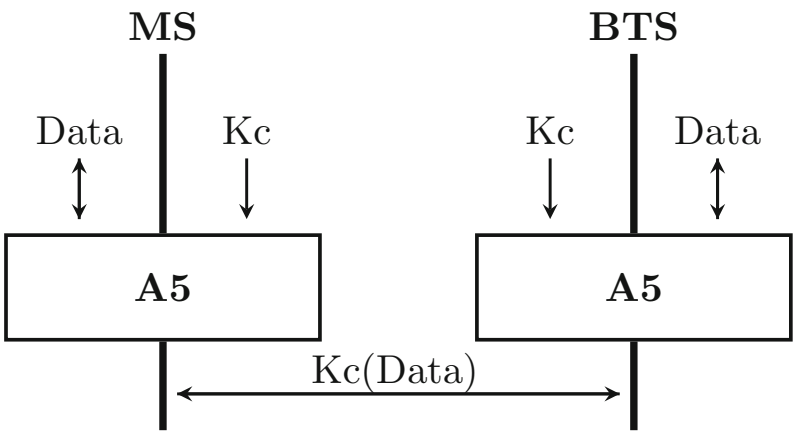


Fig. 4.9 The GSM encryption of user data

AuC sides using the generator algorithm A8, which takes the Ki and the RAND generated in the authentication process as input, as shown in Fig. 4.8. Since both the MS and the network are supposed to share the same Ki and the same RAND, they shall also generate the same Kc on each side. This shared Kc is then used to cipher outgoing data and to decipher received data on both the MS and the BTS sides with the A5 algorithm, which makes a symmetric encryption scheme, as shown in Fig. 4.9.

4.5 Mobility Management

4.5.1 Location Management

As already briefly discussed in Sect. 4.2, an MS must be registered with the PLMN to access GSM services, and such registration is always associated with location management of the subscriber. Generally, a location registration is carried out when the subscriber visits a new network, in which the subscriber reports its current network with its IMSI and receives a TMSI from the VLR, as illustrated in Fig. 4.10 This procedure is initiated by the MS by sending an LU request with its IMSI and current LAI over the RACH to the BSS. This request is then sequentially reported to the MSC, the VLR, and finally the HLR. The HLR then returns a set of authentication parameters (IMSI, Kc, RAND, and SRES) to the VLR, which

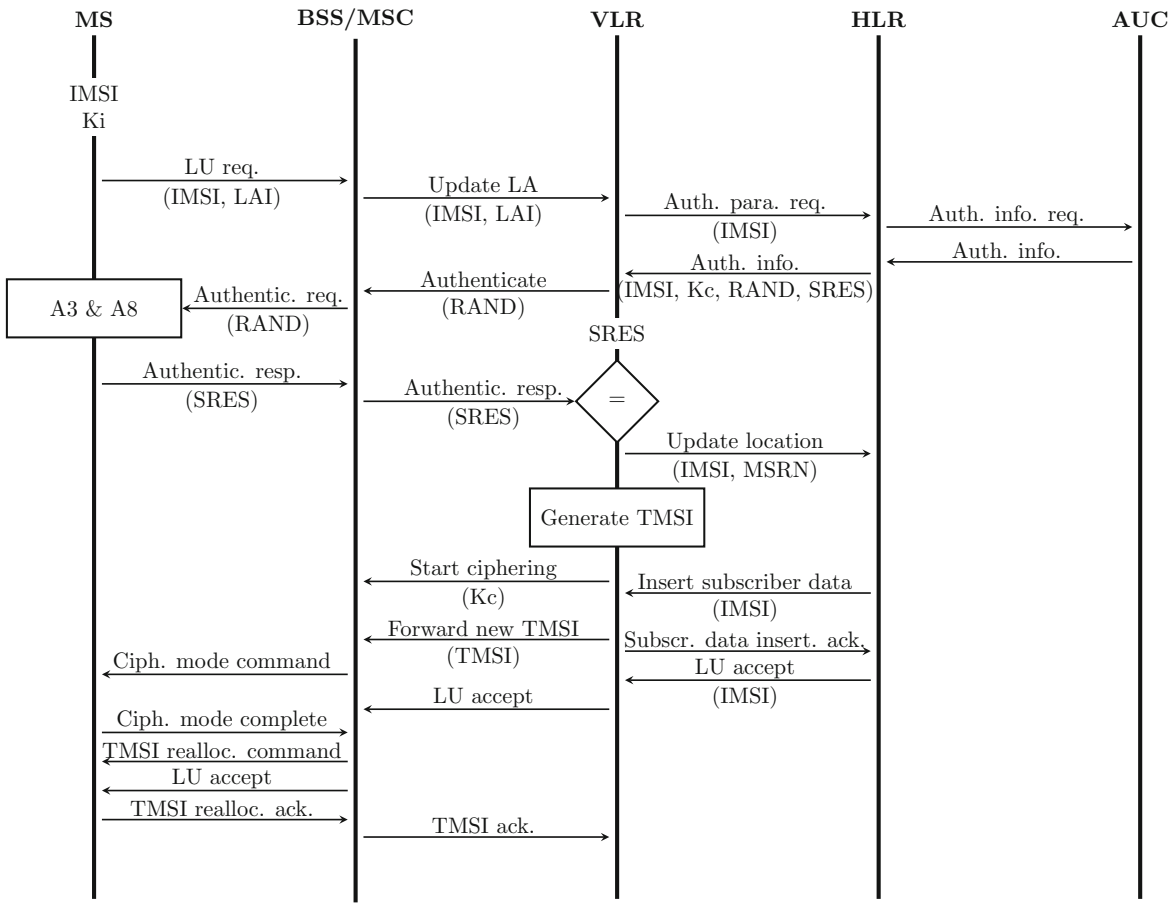


Fig. 4.10 The GSM location registration procedure

forwards only the RAND over the BSS/MSC back to the MS in an authentication request, and keeps the other parameters by itself. The MS then uses the received RAND and its locally stored Ki to generate the Kc and the SRES with the A3 and A8 algorithms, respectively. The generated SRES is then sent back in an authentication response over UL to the VLR to accomplish the authentication process. Once the authentication is passed, the VLR generates a new MSRN and a new TMSI for the subscriber. The MSRN is stored with the LAI in the HLR, while the TMSI is sent back to the MS ciphered. Once successfully allocated to the MS, the TMSI is stored in the SIM as far as the subscriber remains in the same PLMN, so that the location registration procedure does not need to be repeated even after rebooting the MS. Instead, only an LU procedure is triggered in this case.

The LU procedure is also executed, as introduced in Sect. 4.2.1.3, when the MS enters a new LA. It is also initiated by the MS by sending an LU request to the BSS. Differing from the location registration procedure, in LU the MS has already been assigned with its TMSI and key pair. So it sends its TMSI instead of the IMSI with its LU request, which directly triggers an authentication at the VLR. If this LU procedure is executed with change of VLR area, the LU request is first received by the VLR in the new area, which then contacts the old VLR to obtain the authentication parameters. In this case, the new VLR will request the old VLR after a successful authentication to remove the subscriber data, so that an unnecessary storage of redundant data is avoided.

4.5.2 Establishment and Termination of Calls

The establishment of a call follows different procedures, depending on if it is mobile-originated or mobile-terminated.

Mobile-originated calls, as illustrated in Fig. 4.11, always start with the MS sending a setup indication message to the MSC to request a mobility management (MM) connection. Similar to the LU request, the setup indication message is also

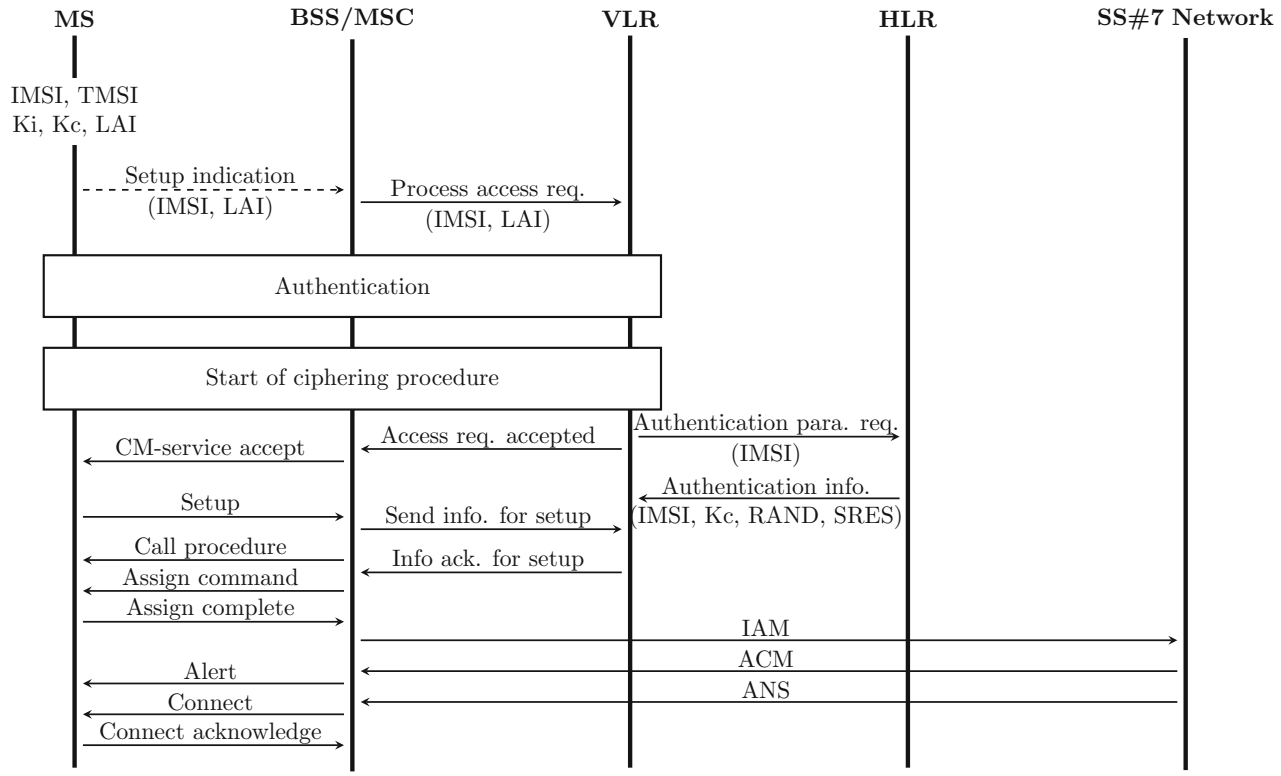


Fig. 4.11 The setup of mobile-originated calls in GSM

sent over the RACH channel and undergoes a random access procedure. Once successfully received by the MSC, the request is then forwarded to the VLR and triggers an authentication process. Upon successful authentication, the ciphering process is started on the air interface, and an encryptedMM connection is established between MS and MSC. The MS then submits the information of its calling target to the MSC for requesting a call connection and gets a conversation channel assigned for it by the MSC. This connection request is signaled to the remote network through the signaling system SS7. The MS is then notified about the call delivery upon answer from the remote exchange and finally a call connection upon answer of the called partner.

Mobile-terminated calls, on the other hand, begin with a message incoming from the SS7 network that arrives at the local MSC, as illustrated in Fig. 4.12. The MSC then contacts the VLR to check the current LAI of the subscriber and initiate a paging procedure in all cells of the corresponding LA. Once successfully received this paging request, the MS sends back a paging response message to initiate the authentication process. After a successful authentication, an MM connection is established so that the MSC can send the call setup information to the MS for confirmation (which generates the call delivery notification that is sent to the caller over the SS7 network). Finally, a call connection will be set up upon answer of the called MS.

The termination of an active call can be triggered either by the MS or by the network. On the air interface, this is executed by means of exchanging three messages: disconnect, release, and release complete. On the network side, the MSC sends to the SS7 network an REL message and receives an RLC message to resolve the connection. Once the connection is terminated, the MSC releases the radio resource and sends to the subscriber's HLR a charging information, as shown in Fig. 4.13.

4.5.3 Handover

As aforementioned in Sect. 3.1, one of the critical issues that GSM was supposed to solve is the reliability of handover.

The first important advance of GSM with respect to its 1G predecessors regarding handover is the introduction of adjacent cell measurements. In 1G system, the handover decisions are made solely based on the signal strength of an MS in its current cell. Once the signal strength falls below a certain threshold, the handover is triggered. This design is ignorant about the signal

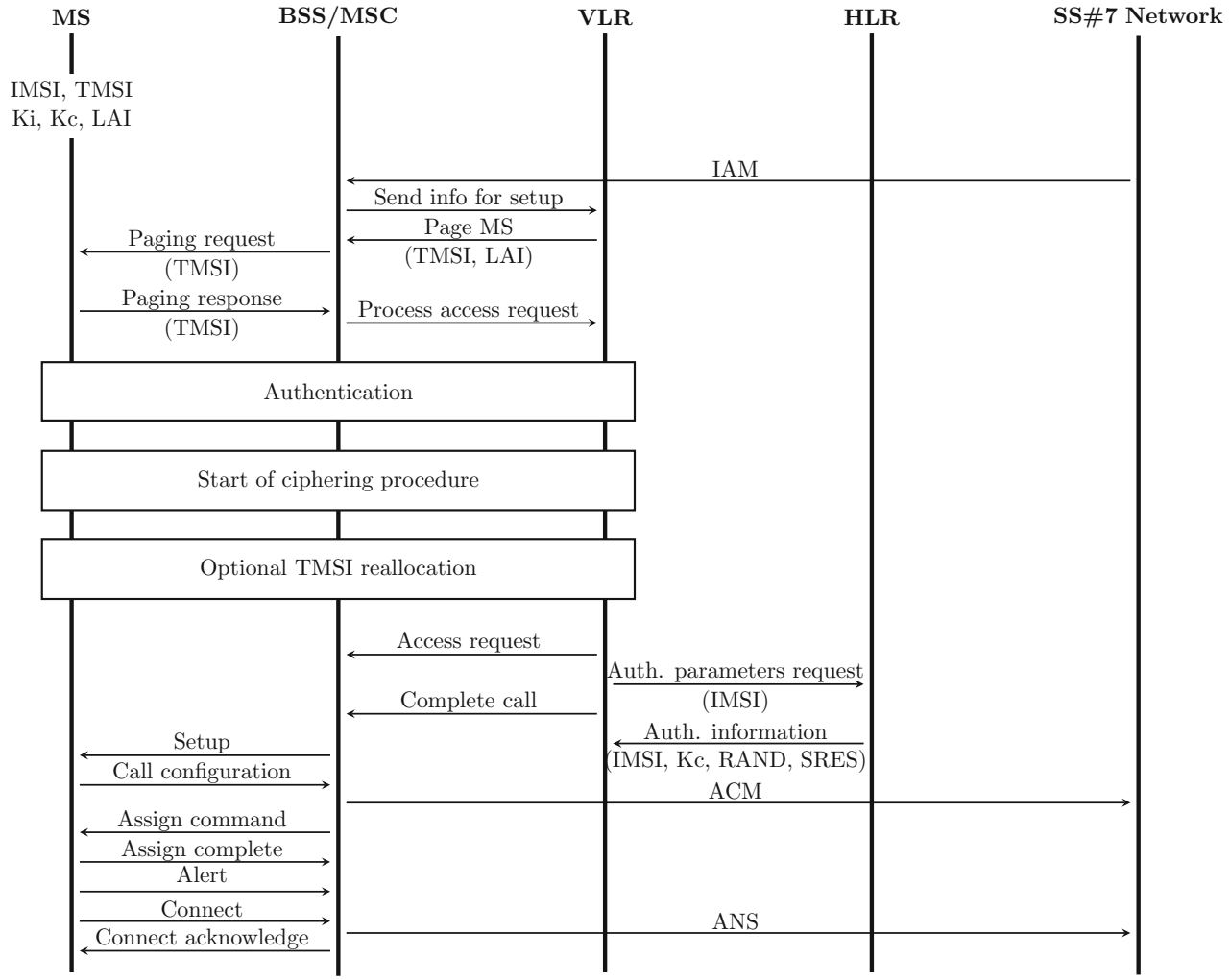


Fig. 4.12 The setup of mobile-terminated calls in GSM

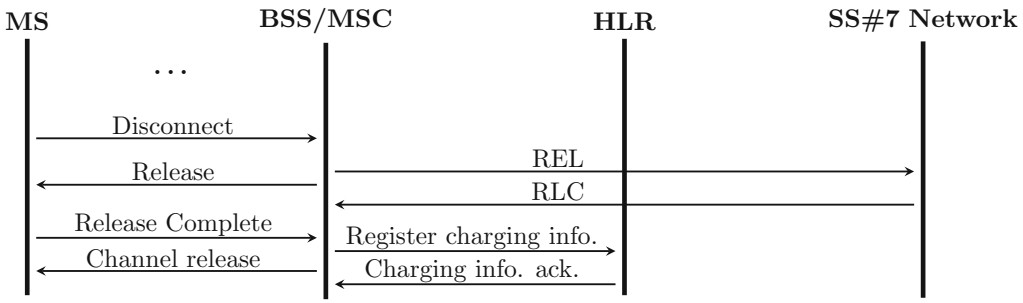


Fig. 4.13 The mobile-initiated call termination in GSM

strength in neighbor cells and may lead to suboptimal handover decisions that cause unnecessary or missed handovers. In GSM, the signal strength is measured not only in the current serving cell but also in its adjacent cells, and the handover decision is made with all adjacent cell measurements taken into account, as briefly illustrated in Fig. 4.14. In GSM Phase 2+ (3GPP GERAN, 2000), the threshold is also dynamically adjusted based on the signal quality.

Another advance introduced later is the mobile-assisted handover mechanism. In the initial phase, GSM followed the 1G design of relying on the base stations to monitor the signal strength, which can be often inaccurate or untimely. In the GSM Phase 2+, this task is assigned to the MSs instead, which significantly enhances the accuracy and timeliness of the channel measurement. The handover decisions are therewith also improved, leading to a better overall call quality.

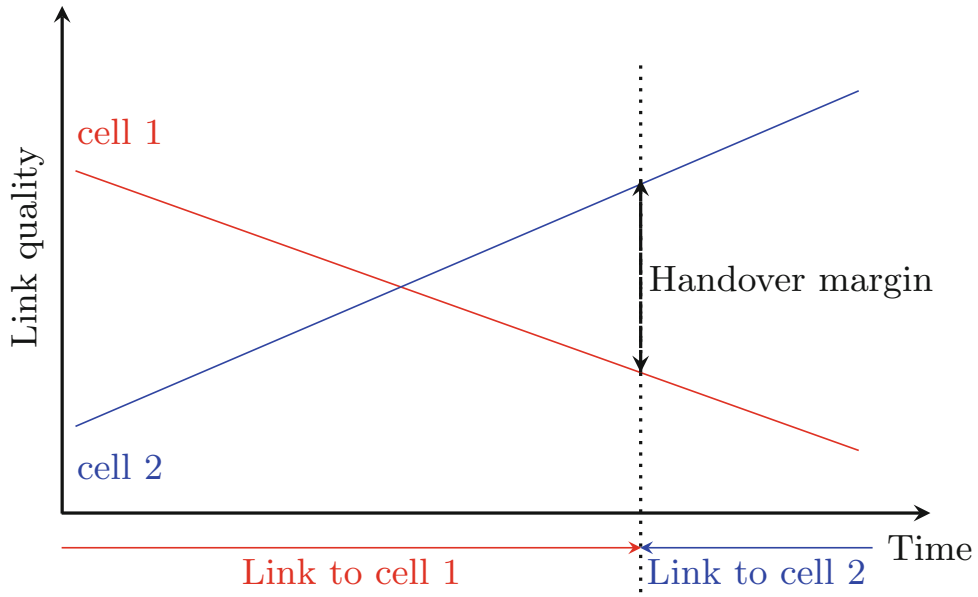


Fig. 4.14 The concept of handover margin in GSM

Moreover, GSM Phase 2+ introduced the so-called soft handover. In the classical hard handover procedure, the old MS first releases its connection to the old BTS and then establishes a connection with the new BTS. In contrast, with soft handover, the old link is released after the establishment of the new link, which improves the quality of calls during the handover at a price of more complex protocol.

Additionally, GSM Phase 2+ also introduced the inter-frequency handover, which allows an MS to maintain its connection while moving between cells that operate on different frequency bands, and therewith enhances the network resource efficiency.

4.6 Summary

This chapter offered an in-depth exploration of the GSM, a 2G mobile communication standard that revolutionized the telecommunications industry. This chapter aims to provide a comprehensive understanding of GSM's architecture, its key technological components, and the challenges it faced during its evolution. We examined the various elements that make up the GSM network and discussed their functionalities in detail. The chapter also covers the protocols and interfaces that enable seamless communication within the GSM network. By looking into the historical context and the regulatory landscape that influenced GSM's development, this chapter provides readers with a nuanced understanding of its impact and legacy. With the proliferation of Internet-based services, along with the rapid technological advancement and widespread availability of mobile electronic devices, the demand for mobile data skyrocketed in the 1990s. This motivated a big leap in cellular networks from being voice-centric to data-centric, adapting to new user behaviors and unprecedented traffic patterns. It resulted in the evolution of cellular technologies toward the 3G, which will be studied in the next chapter.

4.7 Exercises

1. What is the total bandwidth occupied by GSM-900?
2. Please sketch the overall GSM architecture and briefly summarize the functionality of each of its subsystems.
3. What are the identifiers IMEI, IMSI, TMSI, LMSI, and MSISDN? What is the purpose of assigning these multiple identifiers to one subscriber/terminal?
4. Which modulation scheme, multi-access scheme, and duplex mode are used in GSM?
5. How much channel bandwidth does a GSM radio frame occupy? How long is a GSM radio frame in time? How long is a GSM burst in time?

6. How does GSM overcome the frequency-selective fading?
7. Which speech codec is deployed in GSM? What is the coding rate?
8. Which channel coding scheme is deployed in GSM? What is its coding rate?
9. What is the theoretical maximal number of full rate voice calls that can be simultaneously served by a GSM cell? Note: At least one slot per frame is essentially reserved in both UL and DL for signaling.
10. How are PIN and PUK used to protect subscriber information?
11. How is subscriber authentication accomplished in GSM?
12. How is data traffic encrypted in GSM?
13. How does GSM track the location of an MS?
14. What is the criteria to trigger a handover in GSM?

5 Evolution to Third-Generation (3G) Mobile Cellular Communications

5.1 3G: From Voice-Centric to Data-Centric

During the late 1990s and early 2000s, the boom of the Internet resulted in the proliferation of unprecedented services, such as email, web browsing, file sharing, search engines, interactive gaming, e-commerce, multimedia messaging, online high-fidelity music, and video streaming. Global Internet traffic witnessed a phenomenal rise, growing at an average annual rate of 127 percent between 1997 and 2003, from just 5 petabytes per month to a staggering 680 petabytes per month. As mobile electronic devices became more affordable and advanced, these services gradually migrated from wired networks to cellular networks, leading to a significant increase in mobile traffic. Mobile IP traffic experienced a tremendous rise since the middle of the 2000s, as mobile traffic exponentially raised with an average annual growth rate of around 150 percent.

Second-generation digital systems were designed to address the weaknesses, such as limited system capacity, easy eavesdropping, and worse voice quality, in first-generation analog systems. Standards like GSM, IS-95, and IS-136 were designed for optimizing voice communications based on circuit switching networks. Although packet switching sub-networks were integrated into 2G to support data services, the 2.5G standards such as GPRS and EDGE were not capable of providing sufficient data transmission rates and system capacity necessary to support novel high-data rate Internet services. As a result, the need to create a faster and more reliable cellular communication network that could support these new data services drove the advent of 3G technology.

The development of 3G began as early as the initial deployment of 2G, with the aim of creating data-optimized communication systems to replace previous voice-centric cellular networks. The issue of an incompatible mobile environment and fragmented spectrum usage had long been a thorn in the previous generations of mobile telecommunications. The situation called for a global standard that would facilitate full interoperability and interworking. To do so, the ITU-R started working on this initiative in the 1980s. The efforts of the ITU-R eventually culminated in the release of the first recommendation for the Future Public Land Mobile Telecommunications System (FPLMTS) in 1990. The FPLMTS recommendation aimed to resolve the issues associated with the lack of interoperability in mobile communication systems by providing a common set of standards that would allow mobile devices to work seamlessly across different networks worldwide. FPLMTS was renamed to the International Mobile Telecommunications-2000 (IMT-2000) in the late 1990s since the old acronym was hard to pronounce.

IMT-2000 recommendations defined the minimal technical requirements of the 3G system, including high data rate, asymmetric data transmission, global roaming, multiple simultaneous services, improved voice quality, security, and greater capacity. Key features of IMT-2000 include:

- High degree of commonality of design worldwide
- Compatibility of services within IMT-2000 and with the fixed networks
- High quality
- Small terminal for worldwide use
- Worldwide roaming capability
- Capability for multimedia applications, and a wide range of services and terminals

The evaluation criteria specified in ITU-R M.1225 (ITU-R M.1225, 1997) set the target data rates for the 3G circuit-switched and packet-switched data services:

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5_5

- Up to 2 Mbps for an indoor environment
- Up to 144 kbps for outdoor-to-indoor and pedestrian environments
- Up to 64 kbps for a vehicular environment

On the other hand, the ITU, *World Radiocommunication Conference* (WRC) held in February 1992 identified 230 MHz spectrum in the bands 1885-2025 MHz and 2110-2200 MHz for the deployment of IMT-2000 on a worldwide basis. IMT-2000 was an evolutionary standard that aimed to offer high-quality multimedia services, allowing for various access technologies to be used, including satellite and terrestrial cellular systems, thereby providing global roaming capabilities. Based on Radio Regulations provision No. S5.388 and under the provisions of Resolution 212 issued in WRC held in 1997, a.k.a. WRC-97, the bands 1980-2010 MHz and 2170-2200 MHz were assigned for the satellite component of IMT-2000.

One of the phenomenal events in the era of 3G is the auctions of spectrum and licenses in Europe, which left many lessons for the mobile industry (Melody, 2001). The process of allocating the 3G spectrum in Europe began in the late 1990s when the *European Union* (EU) decided to create a single market for wireless communication services across the whole region. This involved allocating spectrum to mobile network operators across the EU member states. Each country was responsible for deciding on the terms and conditions of the allocation process, including the number of licenses to be awarded and the duration of the licenses. The first 3G spectrum assignment in Europe took place in the UK. In 2000, the UK government auctioned 3G licenses, and four mobile network operators won the rights to operate 3G networks: Vodafone, O2, Orange, and T-Mobile. The highest bid was made by Vodafone, which paid 6.1 billion British pounds for its 3G license. The auction raised 22.5 billion pounds for the government, making it the largest spectrum auction in history at the time.

Following the success of the UK auction, many other European countries followed suit and held their own 3G spectrum assignment. The German government auctioned six 3G licenses in the same year, raising a total of 50.8 billion Euros. This was the highest amount raised in any 3G spectrum auction, and it was attributed to the high level of debt that many German mobile operators suffered from. In France, four 3G licenses were issued with a total price of 4 billion Euros in 2001. This was lower than the amounts raised in other countries, but it still represented a significant sum of money for the French government. The four winning bidders were Orange, SFR, Bouygues Telecom, and Free Mobile. The Italian government auctioned five 3G licenses in 2000, raising a total of 11.2 billion Euros, which was a significant amount for a country of Italy's size, and it led to a wave of consolidation within the Italian mobile industry. Four Spanish network operators, namely Telefonica Moviles, Vodafone Spain, Amena, and Xfera, paid a total of 4.4 billion Euros to the Spanish government for four 3G licenses in 2000. In the Netherlands, five Dutch mobile operators, i.e., KPN Mobile, Vodafone Netherlands, Telfort, Ben, and Dutchtone, auctioned five 3G licenses in 2000, with a total cost of 2.7 billion Euros. The Swiss government issued four 3G licenses to Swisscom Mobile, Orange Communications, Sunrise Communications, and diAx, respectively, raising a total of 205 million Swiss Francs. In Sweden, four 3G licenses were issued in 2000 with a total of 7.2 billion Swedish krona. This was a relatively low amount compared to other countries, but it still led to the relatively slow rollout of 3G services in Sweden.

Overall, the cost of 3G licenses varied widely across Europe, with some auctions raising significant amounts of money and others generating more modest revenue. This was a record-breaking amount at the time and exceeded the government's initial estimates by a wide margin. However, the high cost of the 3G spectrum auctions had a significant impact on the mobile industry, as many operators struggled to finance their bids and had to borrow heavily. This led to a wave of consolidation within the industry, with many smaller operators being acquired by larger players.

The high expenditures on 3G licensing resulted in the relatively slow rollout of 3G services in Europe. As a result, the first commercial 3G network was launched by NTT DoCoMo in Japan in October 2001 (Kodama, 2002). The Japanese network, called Freedom of Mobile Multimedia Access (FOMA), operated a new frequency band around 2 GHz and was initially only available in major cities. The first 3G phone released at that time was the NTT DoCoMo FOMA P2101V, which had a color display, video-calling capabilities, and a built-in camera. Other countries soon followed Japan's lead in deploying 3G networks. South Korea launched its first 3G network in 2002, and by 2003, 3G networks were available in many European countries, including the UK, Germany, and Italy. The first 3G network in the US was launched by Verizon Wireless in 2002.

The demand for high-resolution and larger screen displays was boosted by the popularity of video and multimedia content, leading to the evolution of mobile devices. The first smartphone, the IBM Simon, was launched in 1994, while the Nokia 9000, which featured a QWERTY keyboard and combined functions such as email, word processing, and a diary, was released in 1996. On January 7th, 2007, Steve Jobs announced Apple's entry into the mobile phone market. He approached the design of the iPhone by prioritizing its use as a computer rather than just a phone, which revolutionized this industry. The

Apple iPhone proved to be a disruptive piece of technology that redefined mobile phone design and introduced the world to the APP - even though the first model was only a 2G device (Linge & Sutton, 2014).

5.2 IMT-2000 3G Cellular Standards

The ITU played a critical role in developing IMT-2000 in terms of deciding the performance requirements, setting the evaluation criteria and methods, and facilitating international cooperation among industry stakeholders. As an administrative organization, it did not create technical specifications to meet these requirements but only solicited proposals from international or national standardization bodies. Inspired by the successful standardization of GSM, the ETSI initiated a new organization named the 3GPP, which unites other six telecommunications standard development organizations across the world, including

- ARIB (Japan)
- *Alliance for Telecommunications Industry Solutions* (ATIS) (USA)
- *China Communications Standards Association* (CCSA)
- *Telecommunications Standards Development Society, India* (TSDSI)
- *Telecommunications Technology Association* (TTA) (South Korea)
- *Telecommunication Technology Committee* (TTC) (Japan)

The focus of 3GPP was on the standardization of Universal Mobile Telecommunications System (UMTS), also known as WCDMA (Chaudhury et al., 1999). Meanwhile, another group in the USA formed the 3GPP2, aimed at developing the technical specifications for a 3G system as a smooth evolution of the legacy IS-95 standards. The standard developed by 3GPP2 was named CDMA2000, which reused the spectrum bands of IS-95 and inherited the bandwidth setting of 1.25 MHz. Although some differences remained, both 3GPP and 3GPP2 selected CDMA as the underlying baseline technology.

In April 1996, the ITU reached a consensus on a recommendation titled ITU-R M.1225 (1997) - *Guidelines for the assessment of radio transmission technologies for IMT-2000*. The recommendation established several common conditions for evaluating radio transmission technology (RTT), including propagation and multipath fading models, cell deployment models, traffic models, quality of service, and grade of service. The evaluation process focused on four radio environments: indoor, pedestrian, vehicle, and satellite. RTT was assessed using seven criteria, namely spectrum efficiency, coverage efficiency, the impact of technological complexity on installation and operational costs, quality, the flexibility of radio technologies, effects on network interfaces, and the potential for optimizing hand-portable performance.

A circular letter was issued by ITU-R in April 1997, soliciting proposals for a radio system suitable for IMT-2000 from around the world. A total of ten proposals were submitted for terrestrial systems, with eight of them utilizing CDMA technology. Fourteen evaluation groups registered within ITU evaluated these proposals and submitted their evaluation reports in September 1998. At the ITU conference in November 1998, all proposed RTT were recognized to meet the minimum performance criteria for IMT-2000 and were determined as final candidates. In November 1999, the ITU-R released M.1457 Recommendation - *Detailed specifications of the terrestrial radio interfaces of International Mobile Telecommunications-2000 (IMT-2000)* - which recommended five standards as the terrestrial air interfaces for IMT-2000, including

- IMT-2000 CDMA direct spread
- IMT-2000 CDMA multi-carrier
- IMT-2000 CDMA TDD
- IMT-2000 TDMA single-carrier
- IMT-2000 FDMA/TDMA

The ITU-R M.1457 Recommendation was officially approved in May 2000 (ITU-R M.1457, 2000). IMT-2000 CDMA direct spread was developed within 3GPP. This radio interface is called WCDMA, or the Universal Terrestrial Radio Access (UTRA) FDD, where the term UTRA was renamed later as UMTS Terrestrial Radio Access (UTRA). IMT-2000 CDMA multi-carrier was proposed by 3GPP2, commonly referred as to CDMA2000, which consists of the 1X and 3X components. It was a wideband spread-spectrum system that utilizes CDMA in order to meet the needs of 3G, as a smooth evolution of the existing 2G IS-95B standards. The standardization works of IMT-2000 CDMA TDD was led by China Wireless Telecommunication Standards (CWTS). It is also called the UTRA TDD or Time Division Synchronous CDMA (TD-SCDMA). It has been developed with the strong objective of harmonization with UTRA FDD to achieve maximum

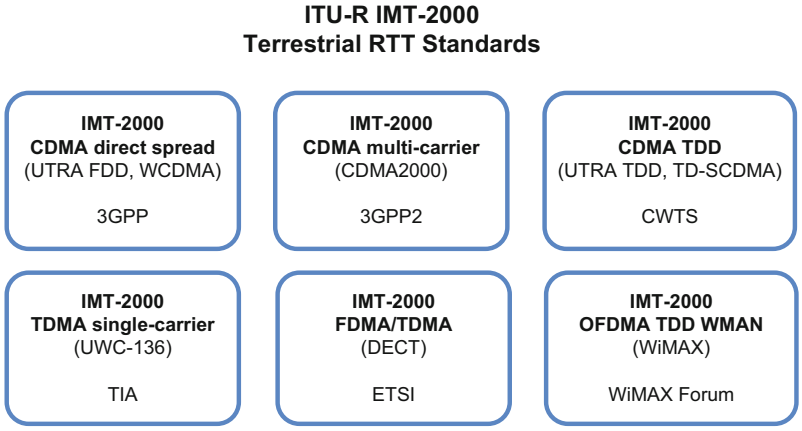


Fig. 5.1 The family of ITU IMT-2000 standards recommended by ITU-R M.1457 (2000)

commonality, where important parameters of the physical layer and a common set of protocols in the higher layers were specified for both FDD and TDD.

IMT-2000 TDMA single-carrier also called Universal Wireless Communications-136 (UWC-136), developed by a consortium consisting of more than 85 wireless network operators and vendors, is based on TDMA for backward compatibility with the IS-136 standard. Digital Enhanced Cordless Telecommunications (DECT) was also called IMT-2000 FDMA/TDMA, which was developed under the DECT Forum and ETSI. Although UWC-136 and DECT were also approved by the ITU-R as 3G standards, they received less support from the industry and were not widely deployed.

In 2007, Worldwide Interoperability for Microwave Access (WiMAX) based on IEEE 802.16 specifications was approved by the ITU-R as the sixth IMT-2000 standard, also known as IMT2000 OFDMA TDD WMAN. Unlike other 3G standards based on CDMA, WiMAX adopted some pre-4G technologies such as OFDM, MIMO, and low-density parity-check (LDPC) codes.

The ITU-R family of IMT-2000 terrestrial radio interface standards is illustrated in Fig. 5.1. In this section, we provide readers with a brief summary of key features of different 3G standards.

5.2.1 Wideband Code-Division Multiple Access/WCDMA

The specifications for IMT-2000 CDMA direct spread/UTRA FDD/WCDMA were developed within 3GPP where the participating standardization bodies include ARIB, ATIS, CCSA, ETSI, TTA, TTC, and TSDSI. The system implements direct-sequence CDMA (DS-CDMA) as its radio access scheme, utilizing a chip rate of 3.84Mchip/s (Mcps) to spread information across a bandwidth of approximately 5 MHz. It was designed to support a variety of services, including circuit-switched services (such as those based on PSTN and ISDN networks) and packet-switched services (e.g., IP-based networks). A flexible radio protocol was designed where several different services such as speech, data, and multimedia can simultaneously be used by a user and multiplexed on a single carrier. The radio bearer supports real-time and non-real-time services by using transparent and non-transparent data transport. The quality of service (QoS) can be adjusted according to metrics such as delay, bit error probability, and frame error ratio.

Initially, WCDMA was designed to operate on the IMT-2000 bands allocated by WRC-92, utilizing 1920 MHz to 1980 MHz as uplink and 2110 MHz to 2170 MHz as downlink frequencies. At WRC-2000 additional spectrum for IMT-2000 was identified and subsequently as a complement to 3GPP Release 99 the relevant specifications have been updated to also include the 3.5 GHz, 2.6 GHz, 1900 MHz, 1800 MHz, 1700 MHz, 1500 MHz, 900 MHz, 850 MHz, and 800 MHz bands as well as a pairing of parts, or whole, of 1710 MHz to 1770 MHz as uplink with whole, or parts, of 2110 MHz to 2170 MHz as downlink.

During the late 1990s, NTT DoCoMo developed Wideband CDMA technology for their own 3G system known as FOMA. WCDMA was selected as the air interface of UMTS, as the 3G successor to GSM. Different systems, including FOMA, UMTS, and J-Phone, shared WCDMA air interface but have different protocols for a complete stack of communication standards. 3GPP submitted it as an IMT-2000 proposal, and the ITU-R approved it as part of the IMT-2000 family standards.

Table 5.1 Third-Generation Cellular Standards

Parameter	WCDMA	CDMA2000	TD-SCDMA	WiMAX
Launch time	2001	2000	2009	2006
Frequency bands	0.8 GHz to 3.5 GHz	0.8 GHz to 3.5 GHz	0.8 GHz to 3.5 GHz	2.3, 2.5, 3.5 GHz
Bandwidth [MHz]	5	1.25	1.6	1.25/5/10/20
Multi-access	CDMA			OFDMA
Duplexing	FDD	FDD	TDD	TDD
Frame length [ms]	10	20	10	5
Modulation	QPSK (DL)/BPSK (UL)		∼8PSK	∼64QAM
Channel coding	Turbo Codes			Turbo/LDPC

NTT DoCoMo launched the first commercial FOMA network in Japan in October 2001, as the successor to i-mode. In Europe, mobile operators had to pay high fees for licenses to use the 3G spectrum, causing significant financial stress. This high licensing cost imposed significant financial pressure on mobile operators, causing a delay in the commercial deployment of European 3G networks.

The network elements of WCDMA are categorized into three main groups: UE, UMTS Terrestrial Radio Access Network (UTRAN) that deals with the radio interfaces, and core network (CN) that is mainly in charge of switching voice calls and routing data packets to and from external networks. In the initial deployment of WCDMA, some CN entities were directly inherited from GSM, aiming to smooth transition and easy deployment. In contrast, the radio part of the legacy GSM network was therefore termed GERAN (Holma & Toskala, 2004). A comparative illustration of the main technical features for major 3G standards are provided in Table 5.1.

5.2.2 Code-Division Multiple Access 2000/CDMA2000

The standardization for IMT-2000 CDMA multi-carrier was conducted within 3GPP2 by a partnership of standardization bodies including ARIB, CCSA, TIA, TTA, and TTC. The physical layer has the capability to handle RF channel bandwidths of $N \times 1.25$ MHz with a chip rate of $N \times 1.2288$Mcps, where N represents the spreading rate number. The specific data rates, channel encoding, and modulation parameters that are available on the traffic channels are defined by radio configurations. Additionally, a versatile multimedia service model is in place that allows any combination of voice, packet data, and circuit data services to be operated. The radio interface also includes a QoS control mechanism to balance the varying QoS requirements of multiple concurrent services.

IS-95 was the first cellular system that employed CDMA technology, making it easy to transition to a CDMA-based 3G standard. When it was upgraded to CDMA2000 as a global IMT-2000 standard, the standardization work was transferred from the TIA to 3GPP2. Being a sister organization of 3GPP, 3GPP2 pushed forward the CDMA2000 technology with an evolution path similar to that of WCMDA. The focus was shifted from circuit-switched voice communications to packet-switched data services. Two parallel evolutionary paths, as demonstrated in Fig. 5.2, were initiated to improve the support of data transmission further. The primary path was *Evolution - Data Only (EV-DO)*, also known as *Evolution - Data Optimized*. The other path aimed to simultaneously support circuit-switched and packet-switched services on the same carrier, hence referred to as *Evolution for integrated Data and Voice (EV-DV)* (Attar et al., 2006).

- **CDMA2000 1x** The initial IMT-2000 CDMA multi-carrier standards approved by the ITU-R had two operational modes: single carrier (CDMA2000 1x) and multiple carriers (CDMA2000 3x). While the 3x mode was crucial for the proposal of CDMA2000 to meet the minimal requirements of IMT-2000, it was never widely deployed in commercial networks. CDMA2000 1x is backward compatible with IS-95/IS-95B, inheriting the basic design of direct-sequence spread spectrum and the channel bandwidth of 1.25 MHz. In addition to several enhancements over the earlier versions of IS-95 to improve spectral efficiency and offer higher data rates, it provided a structure opening the possibility for further evolution of packet-switched data services. CDMA2000 1x could be deployed on IS-95 frequency bands, allowing network operators to upgrade from 2G to 3G without requiring a license for the 3G spectrum. In October 2000, SK Telecom launched the first commercial CDMA2000 1x network in South Korea. As of 2014, the CDMA Development Group reported that 314 operators in 118 countries had provided CDMA2000 1x or 1xEV-DO services.
- **CDMA2000 1x EV-DO Revision 0** The 1x version of CDMA2000 was evolved along different paths, resulting in two options: CDMA2000 1xEV-DV and CDMA2000 1xEV-DO. The former aimed to enhance voice capacity and

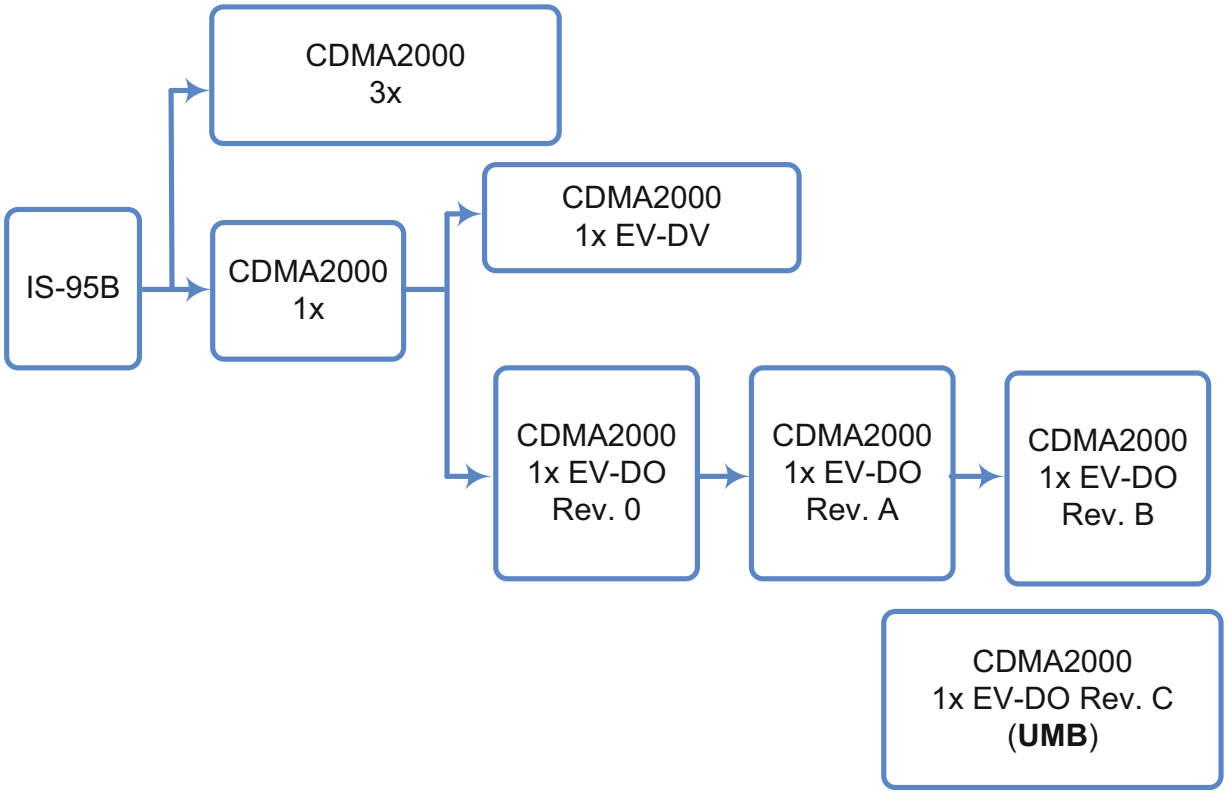


Fig. 5.2 Different revisions of CDMA2000 standards

received limited development under 3GPP2. In contrast, EV-DO was the primary evolution track and underwent several evolutionary steps in Revision 0, Revision A, Revision B, and Revision C. CDMA2000 1x EV-DO was later named High-Rate Packet Data (HRPD). EV-DO Revision 0 redesigned the uplink and downlink structure of CDMA2000 1x, optimized for packet-switched data transmission while removing the constraint of supporting circuit-switched voice communications. An operator would deploy an additional carrier for EV-DO, separating voice and packet data connections on different carriers. In the Revision 0 of CDMA2000 EV-DO, 3GPP2 added a set of data-optimized technologies, including shared channel transmission, channel-dependent scheduling, short transmission time interval (TTI), link adaptation, higher-order modulation (i.e., 16QAM in the downlink), HARQ, virtual soft handover, and receive diversity. Some of these techniques were also adopted by 3GPP for the enhancement of WCDMA. Thanks to a new air interface and a separate channel used solely for data transmission, it can achieve a data rate of 2.4 Mbps in the forward link and 153 kbps on the reverse link over a 1.25 MHz carrier.

- **CDMA2000 1x EV-DO Revision A** The further evolution after Revision 0, named Revision A, focused on enhancing the uplink capacity, similar to HSUPA in 3GPP. Unlike Revision 0, Revision A introduced updates to both the forward and reverse links. The forward link remained similar to Revision 0 but with some improvements, raising the data rate from 2.4 Mbps to 3.1 Mbps. In the reverse link, higher-order modulation (i.e., QPSK and optional support of 8PSK) replaced binary phase shift keying (BPSK) used in Revision 0, and HARQ was introduced, enabling an uplink rate of up to 1.8 Mbps. Furthermore, smaller packet sizes and a shorter TTI were utilized, resulting in a lower latency of up to 50% compared to its predecessor, making it well-suited for supporting voice over Internet protocol (VoIP) and delay-sensitive data services.
- **CDMA2000 1x EV-DO Revision B** The next evolutionary step, Revision B, enabled even higher data rates by utilizing multiple carriers. Up to sixteen carriers could be combined through carrier aggregation to create a 20 MHz bandwidth, theoretically achieving speeds of up to 46.5 Mbps. However, due to hardware and battery life constraints, mobile terminals in a Revision B network typically support up to three carriers, resulting in a maximum peak rate of 9.3 Mbps. The Revision B radio interface was backward compatible with both Revision 0 and Revision A, allowing multi-carrier networks to support legacy single-carrier devices. Additionally, it supported the asymmetric operation, allowing carriers

to be allocated asymmetrically between the downlink and the uplink for applications such as file downloads and video streaming. Moreover, one reverse link could carry control signaling and feedback information for multiple forward links, reducing uplink signaling overhead.

- **CDMA2000 1x EV-DO Revision C** The ultimate version of CDMA2000 1x EV-DO is Revision C, also referred to as Ultra Mobile Broadband (UMB). Unlike the previous revisions, Revision C adopted the Loosely Backward Compatible (LBC) option, which means that it removes the constraint but is not compatible with the previous versions of CDMA2000, facilitating the use of disruptive technologies.

5.2.3 Time Division - Synchronous CDMA/TD-SCDMA

In parallel to the development of UTRA FDD/WCDMA and its evolution to HSPA, 3GPP also worked on the TDD version of UTRA. Regardless of the similar high-layer protocols between FDD and TDD, the physical-layer designs were quite different. There are three chip rate options: the 3.84 Mchip/s TDD option, with information spread over approximately 5 MHz bandwidth, the 7.68 Mchip/s TDD option with information spread over approximately 10 MHz bandwidth, and the 1.28 Mchip/s TDD option, with information spread over approximately 1.6 MHz bandwidth. The UTRA TDD specifications have been developed with the strong objective of harmonization with the FDD component to achieve maximum commonality. This was achieved by harmonization of important parameters of the physical layer and a common set of protocols in the higher layers were specified for both FDD and TDD.

UTRA low chip rate (1.28Mcps) TDD, also called TD-SCDMA (Time Division - Synchronous Code-Division Multiple Access), is substantially different from the other two. The development of TD-SCDMA began in 1998 under the leadership of the Chinese Academy of Telecommunications Technology (CATT) in collaboration with industry partners such as Datang Telecom Technology and Siemens. The goal was to create a 3G mobile communications standard that was domestically developed and could reduce China's reliance on foreign companies for its telecommunications infrastructure. The standard was first proposed by CATT in 2000 and in March 2001, it was approved to merge into Release 4 of the 3GPP specifications as an alternative UTRA TDD version. TD-SCDMA was included in the 3GPP standardization process in the same year and was assigned as one of the 3G standards for China. The significant difference between TD-SCDMA and the other two 3G standards (WCDMA and CDMA2000) is its use of TDD operation instead of FDD for duplex signaling. It adopted a signal bandwidth of 1.6 MHz, 8PSK modulation, and a shorter TTI of 5 ms. Some technical features such as multi-frequency operation and smart antenna/beamforming support with eight antennas were introduced by this system.

Of the three versions, TD-SCDMA was the only UTRA TDD standard deployed on a large scale, with the other two being limited to niche deployment. After years of research and development, TD-SCDMA was commercially launched. China Mobile, the world's biggest mobile operator in terms of the number of subscribers, was granted a 3G license in early 2009 for operating a TD-SCDMA network. It was initially rolled out in eight major cities across the country and was later expanded to cover more than 280 cities. The unique TD-SCDMA deployment worldwide finally became a network consisting of around 500 000 base stations and the peak number of subscribers reached approximately 250 million.

Despite its success in China, TD-SCDMA was not adopted internationally except for small adoption in Pakistan and Nigeria, and the majority of the world's mobile operators chose to deploy WCDMA and CDMA2000. Nonetheless, TD-SCDMA remains a significant part of China's telecommunications history and its development and commercialization serves as an important case study in the country's efforts to build an indigenous technology ecosystem. In addition, it promoted TDD systems' advantages and pushed forward the development of a TDD version of 4G known as TD-LTE or LTE TDD (Chen et al., 2012). The HSPA enhancements of TD-SCDMA were similar to those applied to UTRA FDD, such as the application of high-order modulation (16QAM) and HARQ.

5.2.4 Worldwide Interoperability for Microwave Access/WiMAX

The IEEE 802.16 series of specifications was developed and is maintained by the IEEE 802.16 Working Group on Broadband Wireless Access. These specifications were published by the Standards Association of the *Institute of Electrical and Electronics Engineers* (IEEE) under the umbrella of the Wireless Metropolitan Area Network (WMAN). The first version, released in 2001, was designed for line-of-sight (LoS) communications in the mmWave frequency range of 10–60 GHz, with the goal of providing fixed wireless broadband access. In 2003, the IEEE 802.16a was introduced, which added support for non-line-of-sight (NLoS) operation over the low-frequency bands of 2–11 GHz, but still limited to fixed-wireless-access

applications. The monumental milestone in the series was the release of the IEEE 802.16e-2005 specification in 2005 as the first mobile WiMAX system (Etemad, 2008). This version was empowered by cutting-edge technologies at that time and theoretically offered peak data rates of 128 Mbps in the downlink and 56 Mbps in the uplink over a 20 MHz channel. The first commercial network was deployed in South Korea in 2006, branded as WiBro, and then expanded to many other parts of the world.

IEEE specified a flexible radio interface technology that can be used in various applications, operating frequencies, and regulatory environments. Typically, IEEE 802.16 provides the specifications for the PHY layer and the data link control (DLC) layer instead of the entire protocol stack. The lower element of the DLC layer is the MAC sublayer, and the higher element in the DLC layer is the logical link control (LLC) sublayer. IEEE 802.16 includes multiple physical layer specifications, one of which is known as WirelessMAN-OFDMA. OFDMA TDD WMAN is a special case of WirelessMAN-OFDMA specifying a particular interoperable radio interface. Its PHY layer uses OFDMA and is suitable for a channel allocation of 5 MHz, 7 MHz, 8.75 MHz, or 10 MHz. The MAC uses a connection-oriented protocol designed for point-to-multipoint configuration. It was designed to carry a wide range of packet-switched services, primarily IP-based, while providing precise and instantaneous control of resource allocation for full carrier-class QoS differentiation.

The IEEE 802.16 standards only provide specifications for the PHY and DLC layers, without specifying the overall communication system. On the other hand, these specifications provide multiple alternative options and features, which are not necessary to be implemented in a mobile system. To address this, the WiMAX Forum, a non-profit industry-led alliance, was established to promote and certify compatibility and interoperability of IEEE 802.16-based products. The forum is responsible for selecting technical features from the full set of IEEE 802.16 to create a complete and implementable standard called the WiMAX system profile. The first profile, WiMAX Release 1.0, was published in 2007, with the second profile, Release 1.5, finalized in 2009. In addition, IEEE 802.16e, also known as Mobile WiMAX, was proposed to the ITU-R as a candidate for IMT-2000. In 2007, it was approved by the ITU-R as IMT-2000 OFDMA TDD WMAN, alongside other standards such as WCDMA, CDMA2000, and TD-SCDMA.

The IEEE 802.16 offers several options for the PHY transmission scheme, including OFDM and single-carrier transmission, but Mobile WiMAX is based solely on OFDM transmission. Similarly to LTE, Mobile WiMAX improves spectrum flexibility by adopting variable bandwidths of 1.25, 5, 10, and 20 MHz, achieved by scaling the number of subcarriers (128, 512, 1024, and 2048) under a common inter-subcarrier spacing of 10.94 kHz. IEEE 802.16e specifies both TDD and FDD, including the possibility for half-duplex FDD, whereas the first version of Mobile WiMAX merely supports the TDD operation. In TDD mode, a 5 ms frame is divided into downlink and uplink parts, each consisting of 48 OFDM symbols. Mobile WiMAX supports QPSK, 16QAM, and 64QAM modulation, as well as link adaptation in response to instantaneous channel conditions. IEEE 802.16e supports various channel coding schemes, including Turbo and LDPC codes, similar to HSPA and LTE, but Mobile WiMAX only selected Turbo codes.

5.2.5 Universal Wireless Communication-136/UWC-136

IMT-2000 TDMA single-carrier, also referred to as UWC-136, was developed by TIA in collaboration with the Universal Wireless Communications Consortium, which comprises over 85 wireless network operators and vendors. It utilizes TDMA to provide backward compatibility with the TIA/EIA IS-136 standard, and at the same time, integrate necessary enhancements to meet the IMT-2000 requirements. The main objective was to enable legacy operators with smooth upgrades toward IMT-2000 services, while also offering new operators with competitive features, services, and technology. To evolve IS-136 to IMT-2000, UWC-136 implemented a three-component strategy. Initially, the voice and data capabilities of the 30 kHz channels were enhanced (designated as 136+), followed by the addition of a 200 kHz carrier (EDGE) component for high speed data (384 kbps) to support high-mobility applications (designated as 136HS Outdoor), and a 1.6 MHz carrier component for very high speed data (2 Mbps) in low-mobility applications (designated as 136HS Indoor). The 136HS Outdoor and Indoor components were developed to satisfy the requirements for an IMT-2000 radio transmission technology, with the additional requirement for the consideration of commercially effective evolution and deployment in the existing IS-136 networks.

On the other hand, the development of the IMT-2000 TDMA single-carrier also aimed to achieve maximum commonality with GSM, which had established its dominance in the 2G market. With the identification of additional frequency bands for IMT-2000 at WRC 2000, another evolution path from GSM/GPRS was created by integrating common components

of the IMT-2000 TDMA single-carrier radio interface. Consequently, the later versions of IMT-2000 TDMA single-carrier specifications offer two variations depending on whether an IS-136 circuit-switched network component or a GSM-evolved UMTS circuit-switched network component is utilized. In both scenarios, a common packet-switched network component, namely GPRS or EDGE, can be utilized (Furuskar et al., 1999).

5.2.6 Digital Enhanced Cordless Telecommunications/DECT

IMT-2000 FDMA/TDMA, also called Digital Enhanced Cordless Telecommunications (DECT), was developed under the DECT Forum and ETSI. This radio interface is a general radio access technology for offering a variety of wireless telecommunications from simple residential cordless telephones up to large systems providing a wide range of telecommunications services, including fixed wireless access. It offers high-quality voice and data services, for wide cell radii ranging from a few meters to several kilometers, depending on application and environment.

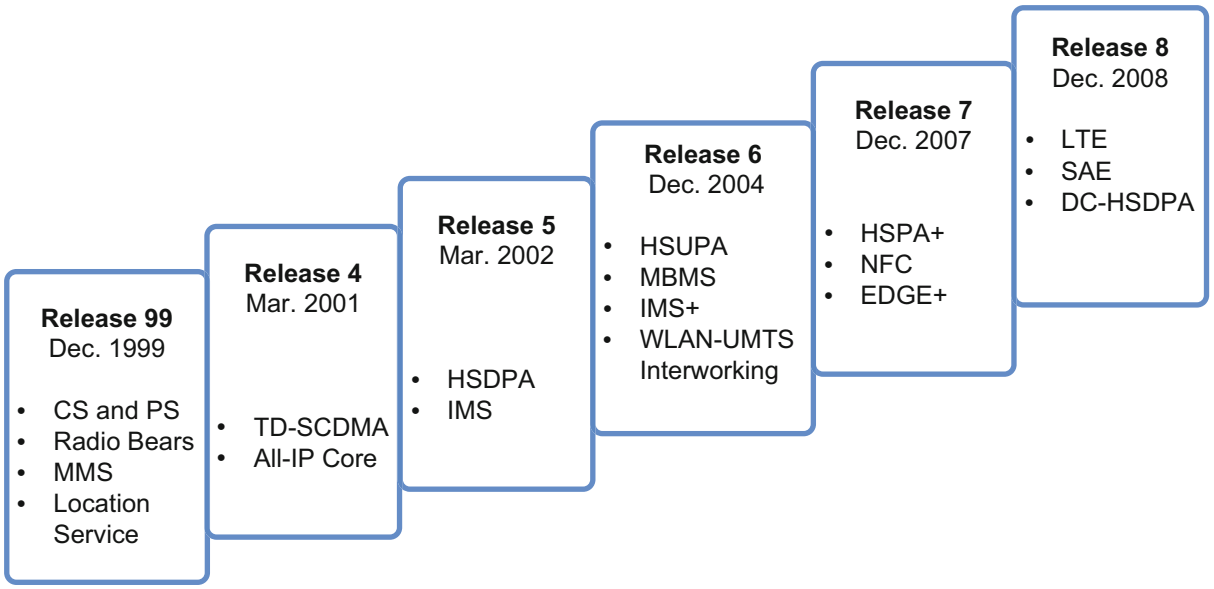
The DECT common interface standard (CI) defines the specifications for the PHY, MAC, DLC, and network layers (ETSI TR101178, 2001). The standard specifies a TDMA radio interface with TDD. The bit rates for the specified modulation schemes are 1.152 Mbps, 2.304 Mbps, and 3.456 Mbps, while higher rates such as 4.608 Mbps and 6.912 Mbps can be supported in the later versions. The CI standard offers comprehensive support for various connection types, including symmetric and asymmetric connections, connection-oriented and connectionless data transport, and variable bit rates up to 2.88 Mbps per carrier. The network layer defines protocols for call control, supplementary services, connection-oriented and connectionless message services, mobility management, and security and confidentiality services. In addition to the CI standard, access profile standards define minimum requirements for accessing specific networks and the interworking of these networks. For example, the generic access profile (GAP) standard defines the requirements when using the speech service, and the DECT packet radio service (DPRS) standard defines the requirements for packet data transport.

5.3 IMT-2000 3.5G Cellular Standards

Since GSM achieved a dominant role in the 2G market, WCDMA, as the evolution of GSM, inherited this advantage and also got a big success in the era of 3G. The initial version of the WCDMA standard, which was specified in Release 99 and Release 4 of the specifications, contains all the technical features to satisfy the IMT-2000 requirements, but its development did not cease there. HSDPA specified in Release 5 was the first significant evolution of the WCDMA radio interface to offer higher data rates. Moreover, the further development of TD-SCDMA was also conducted within 3GPP under the umbrella of HSPA. Therefore, this section presents the main track of WCDMA evolution from HSPA to HSPA+, also taking an eye on the ultimate evolution of CDMA2000, known as UMB.

As illustrated in Fig. 5.3, the evolution of 3GPP WCDMA standards consists of the following milestones:

- *Release 5* enhances the downlink capability to support a rate of up to 14 Mbps, known as HSDPA. This was made possible by the incorporation of several new technical features, such as shared channel transmission, channel-dependent scheduling, higher-order modulation (16QAM), HARQ with soft combining, and link adaptation.
- *Release 6* was finalized in March 2005, which enhanced the uplink air interface known as HSUPA or FDD Enhanced Uplink (EUL), offering a peak rate of 5.76 MHz.
- *Release 7* published in September 2007, as a further evolution referred to as HSPA Evolution or HSPA+. It utilizes multiple antennas technique (2x2 MIMO) and high-order modulation (i.e., 16QAM in the uplink and 64QAM in the downlink), to achieve 28 Mbps in the downlink and 11 Mbps in the uplink over a bandwidth of 5 MHz.
- *Release 8* enables the simultaneous usage of two-layer spatial multiplexing and 64QAM modulation in the downlink. It employed carrier aggregation in a similar way as later done for LTE, thereby leveraging the maximal bandwidth to 10 MHz. Dual-carrier HSDPA can double data rates to 56 Mbps in the downlink by aggregating two carrier channels.
- *Release 9* enhanced the uplink by introducing two aggregated carriers, leading to a rate of 22 Mbps in the uplink.
- *Release 10* can achieve the downlink peak data rate of 168 Mbps by adding the support of aggregating four component carriers for the maximal bandwidth of 20 MHz (Dahlman et al., 2011).



- **CS**: circuit-switched, **PS**: packet-switched, **MMS**: multimedia messaging service, *IMS*: IP multimedia subsystem, **MBMS**: multimedia broadcast multicast service, **NFC**: near field communication, **SAE**: system architecture evolution

Fig. 5.3 Evolution of 3GPP WCDMA standards in releases

5.3.1 High Speed Downlink Packet Access/HSDPA

When Release 99 was finished, HSDPA was not yet being discussed. However, in 2000, it became clear that improvements were necessary to meet the demanding throughput and delay requirements. In order to facilitate this evolution, a study item for downlink enhancement was initiated by 3GPP in March 2000. In March 2002, the 3GPP introduced a new high speed data transfer protocol called *High Speed Downlink Packet Access* to support packet-based multimedia services. It is an evolution of and is compatible with, Release 99 WCDMA systems (Holma & Toskala, 2007).

In order to enable HSDPA operations, the WCDMA system has been enhanced with three new channels: High Speed Downlink Shared Channel (HS-DSCH), Shared Control Channel for HS-DSCH (HS-SCCH), and Dedicated Physical Control Channel (DPCCH). The HS-DSCH serves as the main radio bearer for HSDPA and is provisioned as a shared resource for all users in a specific sector. During each transmission slot, users are assigned an MCS level that maximizes throughput while minimizing the probability of retransmissions. Primary channel multiplexing is done in the time domain, with each TTI consisting of three time slots totaling 2 ms. This short TTI offers several advantages over the 10 ms long TTI used for WCDMA data transfer, including reduced round trip delay and improved channel estimation validity. The HS-SCCH is responsible for signaling information to the UE before the beginning of each scheduled TTI, including the channelization-code set, modulation scheme, transport block size, and HARQ protocol information. The DPCCH is responsible for uplink signaling of Acknowledgements (ACK) and Negative Acknowledgements (NACK) to indicate the status of the previous packet, as well as transmitting channel quality information (CQI) (Holma & Toskala, 2006).

The protocol achieves a substantial increase in throughput through a fast link adaptation scheme that utilizes adaptive modulation and coding (AMC). This means that the protocol transmits at a constant level of power, while dynamically adjusting the MCS to match the current channel conditions experienced by the user equipment. In case of transmission errors, HARQ is used to quickly retransmit packets at the link layer. Fast scheduling played a crucial role in HSDPA, and a significant modification in the new implementation was that the scheduler was situated at Node B. This permits the scheduler to respond swiftly to channel condition changes and guarantees that the UE is serviced even during constructive fading. There are three primary scheduling methods suggested for HSDPA: Round Robin, Maximum Carrier-to-Interference, and Proportional Fair. Soft handoff has been employed by CDMA systems to ensure a smooth handover between base stations. However, the scheduled nature of HS-DSCH renders it impossible to utilize a soft handoff mechanism with HSDPA. Therefore, a rapid, hard handoff algorithm called fast cell selection has been proposed to enable quick switching between base stations. Fast

cell selection functions by monitoring the signal-to-interference ratio (SIR) levels of all base stations in the UE's active set. When a different base station in this set can provide a higher SIR, the user is transferred to the corresponding base station.

5.3.2 High Speed Uplink Packet Access/HSUPA

HSUPA, also referred to as FDD Enhanced Uplink (EUL), was the subsequent advancement of WCDMA after the standardization HSDPA. The main objective of HSUPA was to enhance uplink packet data transmission by achieving data rates of up to 5.76 Mbps, while simultaneously increasing uplink capacity and minimizing latency. The combination of HSDPA and HSUPA is highly beneficial as it optimizes packet data transfer in both the downlink and uplink directions. With both HSDPA and HSUPA active the complete package was called HSPA - High Speed Packet Access (Holma & Toskala, 2006).

During the HSDPA correction phase in September 2002, efforts began to reinforce the uplink for dedicated transport channels via a study item within the 3GPP. Extensive research showed that the investigated techniques provided clear benefits, except that using higher-order modulation in the uplink direction offered no advantages. Consequently, link adaptation was not included in HSUPA. In March 2004, the study item was finalized, and it recommended starting a work item within the 3GPP to specify a HARQ and fast scheduling mechanism for the uplink, along with a shorter uplink TTI. The specifications of HSDPA were first introduced in 3GPP release 6 in December 2004.

In order to enhance uplink data transmission in HSUPA, a new transport channel was introduced called Enhanced Dedicated Channel (E-DCH) (Holma & Toskala, 2006). This implementation created two extra PHY uplink channels, namely, the E-DCH Dedicated Physical Data Channel (E-DPDCH) for user data and the E-DCH Dedicated Physical Control Channel (E-DPCCH) for control information associated with the E-DPDCH. For control purposes of HSUPA, three downlink channels were defined, which include the E-DCH Absolute Grant Channel (E-AGCH) for absolute grants, the E-DCH Relative Grant Channel (E-RGCH) for relative grants, and E-DCH Hybrid ARQ Indicator Channel (E-HICH) for ACK and NACK notification. The E-AGCH is transmitted solely from the serving cell, while the E-RGCH and E-HICH are transmitted from radio links in both serving and non-serving radio link sets.

Uplink scheduling is a crucial aspect of HSUPA, and it is situated in Node B close to the air interface, similar to the downlink scheduler in HSDPA. Its primary purpose is to manage the uplink resources utilized by UEs in the cell by allocating the maximum allowed transmit power ratios to each UE. The scheduling mechanism employs both absolute and relative grants, with the former initializing the scheduling process and providing absolute transmit power ratios to the UE. The latter is utilized for incremental up- or downgrades of the allowed transmit power. To enhance robustness against link adaptation errors, HSUPA utilized the HARQ protocol as a retransmission protocol. In cases where data packets are incorrectly received, Node B can request retransmissions and send either an ACK or NACK to the UE for each packet. Node B can also perform soft combining, which entails combining the retransmissions with the original transmissions in the receiver. HSUPA supports a short TTI of 2 ms, which corresponds to three timeslots, to accelerate packet scheduling and lower latency. Unlike HSDPA, the support of this 2 ms TTI in the UE is not mandatory, but rather a UE capability. During call setup, it is determined whether to use the 2 ms TTI or the 10 ms TTI for HSUPA transmission.

5.3.3 Evolved High Speed Packet Access/HSPA+

To safeguard operator investments in HSPA and facilitate a seamless transition toward LTE, which does not support compatibility with HSPA, a study item on HSPA evolution was initiated in 3GPP in March 2006. Even though initiatives such as MIMO, continuous packet connectivity (CPC), and the "one tunnel" solution for optimizing packet data traffic were already being studied under release 7 in 2005, there was no general plan to guide the evolution of HSPA. Therefore, the effort was established to provide a comprehensive framework for HSPA evolution, referred to as Evolved High Speed Packet Access or HSPA+.

The goal of HSPA+ was to enhance HSPA by providing an incremental evolution path for both the RAN and core network leveraging existing infrastructure. In addition, HSPA+ aims to enable co-existence with the upcoming LTE standards. The guiding principles behind HSPA+ are summarized as follows (Rao et al., 2009):

- Spectrum efficiency, peak data rates, and latency of HSPA+ should be comparable to LTE in a 5 MHz bandwidth.
- The interworking between HSPA+ and LTE should be as smooth as possible and facilitate joint technology operation.

- HSPA+ should be able to operate as a packet-only network, based on the utilization of shared channels only (i.e., HSDPA and HSUPA).
- HSPA+ shall be backward compatible in the sense that legacy terminals compatible with release 99 through release 6 are able to share the same carrier with terminals implementing the latest HSPA+ features, without any performance degradation.

CDMA-based HSDPA in 3GPP release 5 and release 6 already provided an efficient high speed downlink air interface through the use of a short TTI of 2 ms, HARQ, and fast scheduling on a shared channel, facilitated by the use of channel quality feedback and the addition of a new advanced scheduling entity located in the base station. HSDPA offered a peak of 14.4 Mbps. Several enhancements have been introduced for HSDPA in release 7 as part of HSPA+ in order to improve spectral efficiency and cell border throughput. For example, in contrast to QPSK and 16QAM in HSDPA, HSPA+ introduced 64QAM in the downlink, which conveys 6 bits per symbol instead of 4 bits in the case of 16QAM and consequently increases the peak data rate by 50 percent to 21.6 Mbps. HSPA+ utilized closed-loop 2x2 MIMO with two transmit antennas and two receive antennas. Under good channel conditions, dual stream transmissions are possible that can double the peak rate to 28.8 Mbps. HSPA+ employed carrier aggregation in a similar way as later done for LTE, thereby leveraging the maximal bandwidth to 10 MHz or more. 3GPP release 8 specified dual-carrier HSDPA (DC-HSDPA) that can double data rates to 56 Mbps in the downlink by aggregating two carrier channels, in combination with 64QAM and MIMO. Release 10 can achieve the downlink peak data rate of 168 Mbps by adding the support of aggregating four component carriers for the maximal bandwidth of 20 MHz (Dahlman et al., 2011). As a further step, HSPA+ supports eight carriers with peak rates reaching up to 336 Mbps.

The main enhancements for the HSPA+ uplink include the addition of 16QAM in order to improve peak user data rates as well as the CPC feature which allows for an improved always-on connectivity experience. Neither of these features results in significant improvements in uplink spectral efficiency, particularly in the typical macro cellular deployment conditions; hence HSPA+ uplink data capacity still falls short of LTE. 3GPP release 9 enhanced the uplink by introducing two aggregated carriers, leading to a rate of 22 Mbps in the uplink. With the use of MIMO and 64QAM in the later releases, HSPA+ can offer peak rates up to 34.5 Mbps in the uplink.

5.3.4 Ultra Mobile Broadband/UMB

In parallel to the evolution of WCDMA, CDMA2000 was also continuously updated to undergo several evolutionary steps from CDMA2000 1x, 1xEV-DO Revision 0, 1xEV-DO Revision A, to 1xEV-DO Revision B. The ultimate step was CDMA2000 1xEV-DO Revision C, also known as UMB. Adopting the Loosely Backward Compatible option, Revision C is not compatible with the previous revisions of CDMA2000 specifications. The objectives for designing a disruptive air interface were to achieve higher peak rates, improved spectral efficiency, lower latency, and enhanced user experiences for delay-sensitive data applications, like the development of LTE in 3GPP. The significant new features in Revision C are the introduction of typical 4G technologies, namely OFDM and MIMO. In UMB, OFDM multi-carrier transmission uses an inter-subcarrier spacing of 9.6 kHz with different fast Fourier transform (FFT) sizes (128, 256, 512, 1024, and 2048) to flexibly support various transmission bandwidths. Spatial multiplexing supported up to four transmission layers in the forward link but it was only supported in conjunction with OFDM. In the reverse link, up to two spatial layers were specified with codebook-based precoding under the control of a base station. Using a bandwidth of 20 MHz, the maximal rates are 260 Mbps in the forward link and 70 Mbps in the reverse link.

In the initial years of 3G deployment, CDMA2000 and WCDMA were the most potential IMT-2000 standards that competed for the global 3G market. Despite facing challenges such as incompatibility with the legacy GSM standards, late introduction, and high upgrade costs associated with deploying a new air interface technology, WCDMA emerged victorious and eventually became the dominant 3G standard. This resulted in the gradual decline of support for the evolution of CDMA2000, eventually leading to the abandonment of the technology by Qualcomm, the main stakeholder of UMB. In the 3GPP2 camp, UMB was the target to the planned 4G successor of CDMA2000, which competed with LTE in 3GPP. Since 2013, 3GPP2 has been dormant, with no significant activities or developments in the group (Rissen & Soni, 2009).

5.4 Key Technologies for 3G Cellular Systems

The implementation of simultaneous voice and data services during the shift from voice-centric to data-centric communications presented numerous obstacles in designing RAN and core networks. As a result, innovative technologies that could enable advanced multiple access, adaptable resource sharing, variable data rates, and high spectral efficiency were developed. Among a dozen of technical candidates, CDMA, soft handover, the Rake receiver, Turbo codes, AMC, and HARQ, stood out as the fundamental enablers for 3G. This section provides readers with the principles, characteristics, advantages, and challenges of these technologies.

5.4.1 Code-Division Multiple Access/CDMA

In narrowband systems, users within a cell use orthogonal time-frequency resource units through FDMA or TDMA to transmit their signals, while users in adjacent cells are assigned different frequency blocks according to the cellular layout. In contrast, in a CDMA system, each user spreads its information over the entire transmission bandwidth using the direct-sequence spread spectrum (DSSS) technique by multiplexing a pseudo-random sequence (Pickholtz et al., 1982), known as the spreading code. The transmission bandwidth in CDMA is much larger than the original signal bandwidth, and their ratio is referred to as the *processing gain*. For instance, IS-95 is an example of a spread-spectrum system with a chip rate of 1.2288 MHz and a typical data rate of 9.6 kHz, amounting to a processing gain of 128. In spread-spectrum systems, since the symbol rate per user is low, ISI is generally negligible, and equalization is simplified. The Rake receiver, which is a simpler receiver, can be used to extract frequency diversity.

CDMA is distinct from narrowband system design since all users share all available time-frequency resources with the price of inter-user interference. The large processing gain of each user helps mitigate interference from other users, which is perceived as random noise. As a result, the system is *interference-limited* rather than *degree-of-freedom-limited*. With orthogonal spreading codes such as Walsh-Hadamard codes, the information signals of different users can be well separated. Non-orthogonal spreading codes are also effective to differentiate multi-user signals, but mutual interference raises. In addition to enabling multiple access and providing frequency diversity against multipath fading, CDMA offers other benefits, including anti-jamming capabilities against intentional interferers and enhanced privacy in the presence of eavesdroppers (Tse & Viswanath, 2005).

Compared with narrowband systems, the CDMA cellular network has the following advantages:

- **Universal Frequency Reuse -** In FDMA or TDMA systems, inter-user interference is avoided by assigning disjoint time-frequency slots to different users within the same cell and non-overlapping frequency bands to adjacent cells. The number of degrees of freedom per user is significantly reduced in terms of the number of users per cell and the frequency reuse factor. However, in a CDMA system, not only do users in the same cell share the same time-frequency resources, but users in different cells, including adjacent ones, also share the same resources. This approach increases the degrees of freedom per user and simplifies network planning. It achieves maximal utilization of transmission resources, resulting in a frequency reuse factor of 1, referred to as *universal frequency reuse*. These benefits come at the expense of a lower signal-to-interference-plus-noise ratio (SINR) for individual links.
- **Soft Capacity -** In a narrowband system, the time-frequency resources are divided into a fixed number of orthogonal channels. This imposes a *hard capacity* limit on the system since new users cannot be admitted into a network once the time–frequency slots run out. In contrast, a CDMA system is scalable, where the number of users can be increased flexibly only with the cost of the increasing level of interference. This allows more graceful degradation of the performance of a system and provides *soft capacity* on the system.
- **Interference Sharing -** In FDMA or TDMA systems, an allocated channel remains idle during the silence period of a voice conversation when the user is listening and cannot be shared with other users. It is inefficient use of resources. However, in a CDMA system, the capacity is interference-limited and there is no hard limit on the number of channels. This means that a CDMA system can automatically benefit from the variability in user activity. Since a user's performance depends solely on the aggregate interference level, all other users benefit from the decreasing interference level when a user stops transmitting data.

- **Soft Handover -** A user on the edge of a cell in a CDMA system can receive or transmit signals to two or more base stations due to the common spectrum shared by all cells. This technique is called soft handoff and is another diversity technique but at the network level (sometimes called macro-diversity). Soft handoff is an important mechanism to increase the performance of CDMA systems.

To ensure reliable performance, accurate power control is crucial in a CDMA system, particularly in the uplink, due to *the near-far effect*. Even if two users transmit at the same power, the signal from the nearby user can be tens of decibels stronger than that from the user at the cell edge. The far user may experience difficulty in detecting its desired signal amidst the overwhelming inter-user interference. Therefore, power control is necessary to adjust the power of each user's signal, so that all users receive a signal strength that is roughly the same. In addition, power control makes sure that there is no strong inter-cell interference since all cells share the same spectrum. However, the real-time measurement of signal strength and frequent transmission of power control signaling introduce significant overhead. In contrast, power control is optional in narrowband systems and is primarily used to reduce power consumption, rather than to mitigate interference.

5.4.2 Soft Handover

The neighboring cells in narrowband systems are assigned to different frequency blocks and the mobile terminals can only tune to a single carrier frequency at a time. Consequently, a terminal has to disconnect the outgoing cell before connecting to a new cell. Such a *hard handover* causes the drop of voice calls, which was one of the major problems worsening the user experience in the first- and second-generation systems. Since the neighboring cells use the same frequency in a CDMA system, a user at the edge of a cell can communicate with more than one base station simultaneously. In the uplink, the user signal is modulated by multiple spreading codes associated with different cells and then is received by more than one base station. The received signals are combined and detected at a central processing unit such as the radio network controller (RNC) in WCDMA or MSC in GSM. In the downlink, the same signal can be transmitted to the user by more than one base station simultaneously. Soft handover is also a diversity technique sometimes called macro-diversity (Wong & Lim, 1997).

Soft handover is a mobile-initiated process that involves a user tracking the downlink pilot of the serving cell and searching for pilots of adjacent cells, which are known pseudo-noise sequences shifted by offsets. When a pilot is detected and found to have sufficient signal strength relative to the pilot from the serving cell, the mobile station informs the event. The serving base station notifies the switching center, which enables the target cell's base station to simultaneously send and receive the same traffic to and from the mobile station. In the uplink, each base station demodulates and decodes the frame or packet independently, and it is up to the switching center to arbitrate. Normally, the better cell's decision will be used. Soft handover provides a form of receive diversity by viewing the base stations as multiple receive antennas. Although maximum-ratio combining (MRC) is the optimal processing of signals from multiple antennas, it is difficult to do in the handover scenario as the antennas are geographically apart. Instead, soft handover achieves *selection combining*. There is another form of handover called *softer handover*, which takes place between sectors of the same cell. In this case, since the signal from the mobile is received at the sectored antennas, which are co-located at the same base station, MRC can be performed.

5.4.3 Rake Receiver

Multipath fading is the major impairment to limit the performance of wireless transmission, where several delayed copies of the transmitted signal, referred to as multipath components, combine at the receiver. This combination is sometimes constructive but usually destructive. Since each multipath component contains the original information, the ideal case is to add all multipath components coherently to improve the power of the desired signal and lower the probability of deep fades. To this end, an advanced receiver structure called the Rake receiver can be employed. It has several branches, referred to as fingers, to deal with different signal paths. Each finger is synchronized to a signal path and equipped with an individual correlator to despread the received signal. The outputs of the Rake receiver's fingers are combined to detect the transmitted symbols, where different methods such as selection combination and MRC can be applied. Direct-sequence CDMA is well-suited for the application of Rake reception since the large signal bandwidth allows a high resolution of distinguishing multiple paths. It is particularly suitable for WCDMA transceiver with a signal bandwidth of 5 MHz. The Rake reception is in essence a diversity technique, which was proposed by Price and Green (1958). It has been described as the historically most important adaptive receiver for multipath fading channels.

5.4.4 Turbo Codes

According to information theory, a code chosen randomly with a sufficiently large block length can approach the Shannon capacity. However, the complexity of the maximum-likelihood decoding of such a code increases exponentially with the block length, making decoding physically impractical. In 1993, Berrou, Glavieux, and Thitimajshima invented a powerful coding scheme that can transmit data with a code rate within a fraction of a decibel away from the Shannon capacity on a Gaussian channel (Berrou et al., 1993). This coding scheme combines parallel concatenation of convolutional codes, large block lengths, interleaving, and iterative decoding while relying on the exchange of soft-decision information. This innovative coding scheme is a significant advancement in the field of information theory and enables more efficient and effective data transmission.

A typical Turbo encoder consists of two parallel convolutional encoders and an interleaver, as shown in Fig. 5.4. The information bits m concatenate with two parity-bit sequences X_1 and X_2. One convolutional encoder generates X_1 in terms of m, while another convolutional encoder makes X_2 using the interleaved information bits. The concatenated data (m, X_1, X_2) is modulated and then transmitted to the receiver. The iterative decoding of Turbo codes is implemented by two parallel decoders, an interleaver, and a deinterleaver. Berrou et al. used maximum a posteriori probability (MAP) algorithm to perform maximum-likelihood (ML) estimation, yielding reliable information (soft decision). To be specific, the first decoder generates

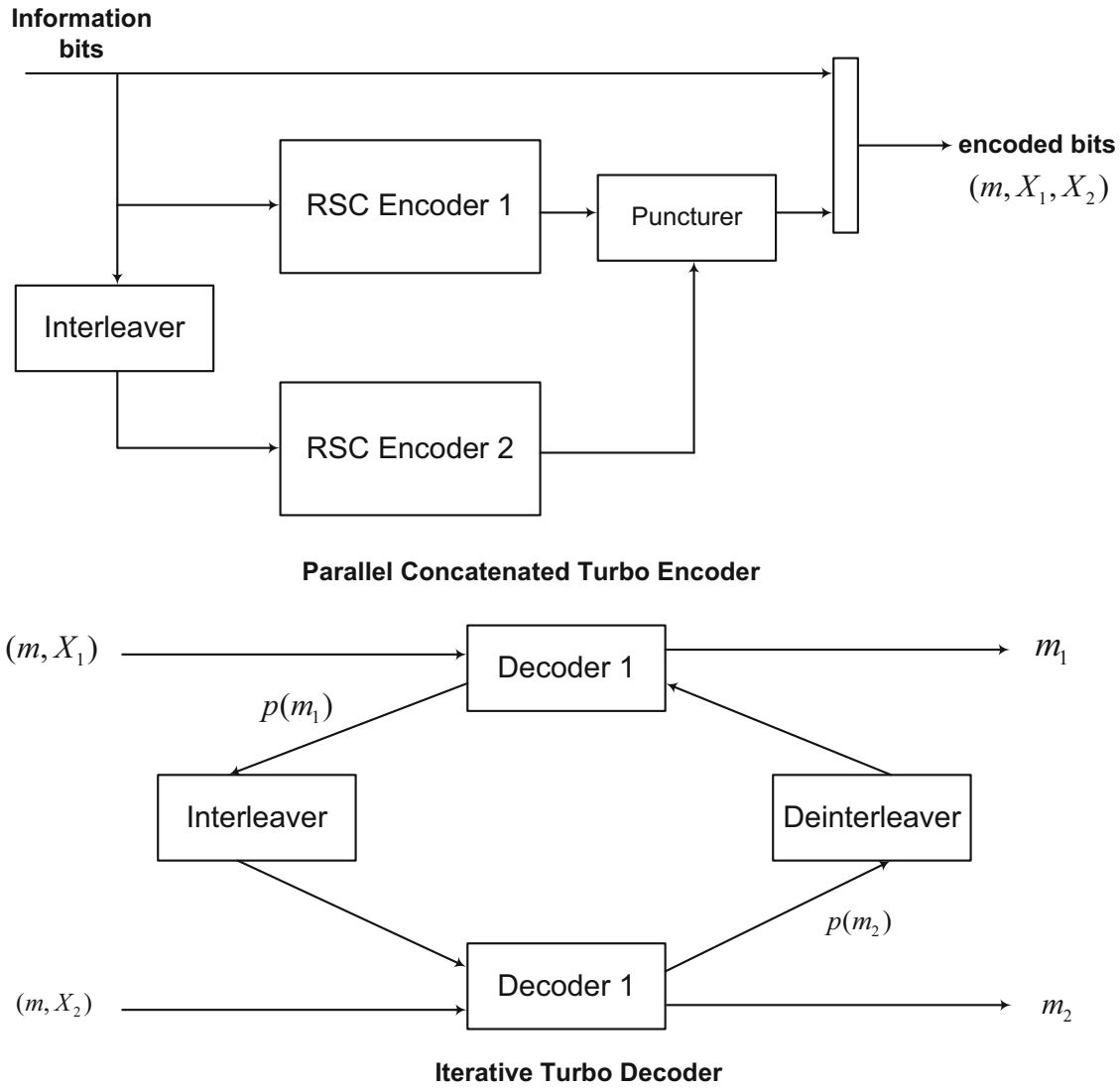


Fig. 5.4 The basic structure of Turbo encoder and decoder

a probability measure $p(m_1)$ on each information bits based on the codeword (m, X_1). This probability measure is passed to the second decoder. Then, another probability measure $p(m_2)$ is generated by the second decoder in terms of (m, X_2) and is passed to the first encoder. This process iterates until the convergence condition reaches when, ideally, the decoders eventually agree on the probability measure.

Berrou et al. exhibited that a bit error rate (BER) of 10^{-5} can be achieved for an E_b/N_0 at 0.5dB above the Shannon capacity for 1/2-rate QPSK on a Gaussian channel, and this gap was narrowed to 0.35dB later (Berrou & Glavieux, 1996). Turbo codes received wide recognition quickly and triggered a wave of research and development addressing its design, implementation, performance evaluation, and application in digital communications systems. The CDMA2000 system employed 1/5-rate Turbo coding that consists of two identical, eight-state, parallel, 1/3-rate Recursive Systematic Convolutional (RSC) encoders. This code is punctured to rates between 1/2 and 1/4, with the interleaver length ranges from 378 to 12,282. Channel coding adopted by WCDMA is similar to that of CDMA2000, except that two Turbo encoders use different pseudo-random interleavers.

5.4.5 Adaptive Modulation and Coding/AMC

Adaptive modulation and coding or AMC is a technique that allows for reliable and efficient transmission over channels that vary over time. This approach involves the receiver estimating the channel and sending this information back to the transmitter, which then adjusts the transmission scheme based on the characteristics of the channel. Traditional modulation and coding techniques that do not adapt to changing channel conditions require a fixed link margin to ensure acceptable performance even when the channel quality is poor, making them inefficient in utilizing the channel. This is because Rayleigh fading can cause a significant power loss of up to 30 dB. However, by adapting to the channel fading, AMC can increase the average throughput, reduce the required transmit power, or lower the average probability of bit error by utilizing favorable channel conditions to transmit at higher data rates or lower power, and adjusting the data rate or power as the channel quality degrades. There are several practical factors that influence the suitability of AMC. To enable adaptive transmission, a feedback path between the transmitter and receiver is necessary, but this may not always be possible in some systems. Additionally, adaptive techniques may perform poorly if the channel changes too quickly for reliable estimation and feedback to the transmitter. Wireless channels often exhibit fluctuations on different timescales, namely fast fading and slow fading. In many cases, only the slower variations can be tracked and adapted to, requiring flat fading mitigation to address the effects of multipath (Goldsmith, 2005).

5.4.6 Hybrid Automatic Repeat Request/HARQ

HARQ is a method of ensuring the reliable transmission of data by combining forward error correction (FEC) with ARQ. Unlike the conventional ARQ, which adds parity bits using an error-detecting code like CRC directly over the information bits, HARQ encodes the original data with an FEC code first and then adds parity bits. If a corrupted message is detected but cannot be corrected by FEC, the HARQ operation requests retransmission. HARQ with soft combining stores the erroneously received packet in buffer memory and combines it with the retransmission for a more reliable packet.

In any HARQ system, retransmission must contain the same set of information bits as the original transmission. However, the set of coded bits sent during each retransmission can be selected differently, as long as they represent the same set of information bits. HARQ with soft combining is typically classified into two categories: *Chase combining* and *incremental redundancy*.

- *Chase combining* requires the retransmitted bits to be identical to the original transmission, which was first introduced by Chase (1985). In this technique, the receiver uses the MRC to merge each received message with previous transmissions. The combined coded bits are then delivered to the decoder. Since each retransmission is an exact copy of the original transmission, using Chase combining is akin to adding more repetition coding. Consequently, this technique does not provide any additional coding gain because no new redundancy is transmitted. However, it does increase the received energy per bit to accumulate a higher SNR for each retransmission.
- *Incremental redundancy* does not require each retransmission to be a copy of the original transmission. Instead, multiple sets of coded bits are generated, each representing the same set of information bits. When retransmission is necessary, it generally employs a different set of coded bits than the previous transmission. The receiver then merges the retransmission with previous transmission attempts of the same packet. As the retransmission may have additional parity bits that were

not included in previous transmission attempts, the resulting code rate is typically reduced. Moreover, the number of coded bits in each retransmission need not be the same as in the original transmission, and different modulation schemes can be employed for different retransmissions. Consequently, incremental redundancy can be seen as a more general form of Chase combining, or conversely, Chase combining can be viewed as a special case of incremental redundancy.

5.5 Summary

Internet-based services, along with the rapid technological advancement and widespread availability of mobile electronic devices, motivated a big leap in cellular networks from being voice-centric to data-centric. A disruptive modulation method, i.e., CDMA, was exploited to power the successful evolution of 3G cellular systems. This chapter delved into the major driving forces behind the transition from 2G to 3G, summarized the main technical standards that constitute 3G, including WCDMA, CDMA2000, TD-SCDMA, and WiMAX. Then, we explored the fundamentals of CDMA that propelled the evolution of 3G, along with other key 3G technologies. The next chapter will introduce the 3G successor standard of GSM: the UMTS. We will provide a comprehensive overview of UMTS regarding its key features, architecture, radio interface, security, mobility management, location service, and the IP Multimedia Subsystem (IMS), based on 3GPP Release 7.

5.6 Exercises

1. Describe the background of developing 3G cellular technology in the 1990s.
2. Can you identify the phenomenal event in the era of 3G that never happened in other generations of mobile communications? What lessons can learn from this issue?
3. The target data rates for the 3G circuit-switched and packet-switched data services are
   - Up to 2 Mbps for an indoor environment
   - Up to 144 kbps for outdoor-to-indoor and pedestrian environments
   - Up to 64 kbps for a vehicular environment

   Do you know which organization is responsible for specifying these performance indicators? Find the document that provides these recommendations.
4. Select the multiple access technologies adopted by 3G cellular standards:
   a. Frequency-division multiple access (FDMA)
   b. Time-division multiple access (TDMA)
   c. Code-division multiple access (CDMA)
   d. Orthogonal frequency-division multiple access (OFDMA)
5. In November 1999, the ITU-R released M.1457 – *Detailed specifications of the terrestrial radio interfaces of International Mobile Telecommunications-2000 (IMT-2000)* – which recommended five standards as the terrestrial air interfaces for IMT-2000, including
   - IMT-2000 CDMA direct spread
   - IMT-2000 CDMA multi-carrier
   - IMT-2000 CDMA TDD
   - IMT-2000 TDMA single-carrier
   - IMT-2000 FDMA/TDMA

   In 2007, Worldwide Interoperability for Microwave Access (WiMAX) based on IEEE 802.16 specifications was approved by the ITU-R as the sixth IMT-2000 standard, also known as IMT2000 OFDMA TDD WMAN.
   Identify the most successful standard in the commercial market. What were the main reasons making this standard win?
6. Which pre-4G technologies were first employed in WiMAX systems?
7. What are the main technical features enabled by CDMA?
8. What does universal frequency reuse mean? How can CDMA achieve this?
9. If we divide the time-frequency resources into a fixed number of orthogonal channels, the maximal number of users can be supported simultaneously is known. It is called a *hard capacity* limit on the system since new users cannot be admitted into a network once the time–frequency slots run out. In contrast, a CDMA system is scalable, where the number of users can be increased flexibly. Describe the cost of achieving this *soft capacity*.
10. Describe the difference between *hard handover* and *soft handover*.

Universal Mobile Telecommunications Service (UMTS)

6

6.1 Frequency Bands and Key Features

The RAN of UMTS, known as UTRAN, has different specifications upon the deployment region. The duplex mode, and therefore the frequency allocation as well, vary with the standard. The most widely deployed standard of UTRA is the WCDMA that uses FDD, with the operation bands listed in Table 6.1. In some countries, primarily China, two TDD versions of UTRA are developed and standardized, namely the Time Division CDMA (TD-CDMA) and TD-SCDMA. The former was only shortly deployed during the early 2020s before being replaced by the latter and other successor technologies. For UTRA-TDD, the frequency bands listed in Table 6.2 are specified.

Compared to 2G systems, UMTS exhibits the following main new features:

- **CDMA technology**: UMTS uses CDMA, which allows for greater capacity and improved spectrum efficiency compared to GSM/GPRS/EDGE.
- **Packet-switched core network**: UMTS uses a packet-switched core network, which allows for more efficient use of network resources and faster data transfer rates.
- **High speed data transfer rates**: In Release 99, UMTS supports high speed data transfer rates up to 384 kbps. In Release 7, this maximal rate is raised to 28 Mbps in DL and 11.5 Mbps in UL.
- **Increased network capacity**: UMTS provides increased network capacity through the use of wider frequency bands and more efficient use of available spectrum.
- **Multimedia services**: UMTS provides support for multimedia services such as video conferencing, streaming video, and high speed Internet access.
- **Enhanced voice services**: UMTS provides enhanced voice services such as better call quality, improved voice clarity, and reduced background noise.
- **Increased security**: UMTS provides increased security features, such as user authentication and encryption, to protect against unauthorized access and eavesdropping.
- **Advanced roaming capabilities**: UMTS supports advanced roaming capabilities, allowing users to access services and features seamlessly while roaming between different networks and countries.
- **Support for location-based services (LBS)**: UMTS provides support for LBS, allowing operators to offer services based on the user's location, such as location-based advertising, tracking, and emergency services.

Albeit there are FDD (WCDMA) and TDD TD-CDMA/TD-SCDMA versions of UTRA, they are not capable of coexisting in the same UTRAN and must be therefore separately specified. In this chapter, we only focus on WCDMA to explain the design of UMTS. Readers with particular interest in UTRA-TDD are referred to Esmailzadeh and Nakagawa (2003).

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5_6

Table 6.1 The paired UTRA–FDD bands (3GPP TS 25.101, 2010)

Operating band	UL frequencies (MHz)	DL frequencies (MHz)	Tx-Rx frequency separation (MHz)
I	1920–1980	2110–2170	190
II	1850–1910	1930–1990	800
III	1710–1785	1805–1880	95
IV	1710–1755	2110–2155	400
V	824–849	869–894	45
VI	830–840	875–885	45
VII	2500–2570	2620–2690	120
VIII	880–915	925–960	45
IX	1749.9–1784.9	1844.9–1879.9	95
X	1710–1770	2110–2170	400

Table 6.2 The UTRA–TDD bands (3GPP TS 25.102, 2011)

Operating band	Frequencies (MHz)	Note
A (lower)	1900–1920	
A (upper)	2010–2025	
B (lower)	1850–1910	Used in ITU Region 2
B (upper)	1930–1990	
C	1910–1930	Used in ITU Region 2
D	2570–2620	Used in ITU Region 1

6.2 System Architecture and Interfaces

The overall system architecture of UMTS can be briefly illustrated by Fig. 6.1. It is divided at the highest level into three major subsystems: the UE, the UTRAN, and the CN.

The term UE refers to mobile devices in UMTS and its successor systems and is distinguished from the term MS in GSM for its support to advanced functionalities and capabilities that are not provided by GSM, GPRS, or EDGE networks. Similar to MS, a UE also consists of an ME and an identification module named Universal SIM (USIM), which are interconnected to each other through the Cu interface. Compared to SIM, USIM provides enhanced security and authentication features, in addition to support to multimedia services.

The UTRAN has a similar topology to GERAN, with the so-called Node B to replace the BTSs and the RNCs to replace the BSCs. The RNC manages its subordinate Node B over the Iub interface, and the Node B is connected to the UEs over the air interface known as Uu interface. Note that UMTS was designed to be backward compatible with GSM/GPRS/EDGE networks, so that a UE can also connect to a GERAN-BTS over the Um interface, and a GSM MS can also access the UTRAN via the Uu interface.

The UMTS CN is responsible for managing and routing the data and voice traffic of UMTS and compatible with the 3GPP 2G core network architecture. It takes over the elements of MSC, VLR, GMSC, and EIR from GSM and SGSN as well as GGSN from GPRS. Especially, UMTS introduces the Home Subscriber Server (HSS) to replace HLR and AuC, combining their functionalities together and providing a centralized database for subscriber information. The UMTS CN is connected to UTRAN and GERAN at RNC and BSC, respectively, via two parallel interfaces: the IuCS at MSC for circuit-switched voice service and the IuPS at SGSN for packet-switched data service. Furthermore, UMTS supports the 3GPP-standardized core network architecture of IMS on top of its core network, in order to deliver IP-based multimedia services as a key feature of it. The IMS will be introduced later in Sect. 6.8.

6.3 Physical Layer

6.3.1 Physical Channels, Channel Mapping, and Radio Frame

To understand the physical layer design of UMTS, it is necessary to have an overview to the definition of its physical, transport, and logical channels.

Generally, there are six types of logical channels in UMTS, namely the BCCH, the Paging Control Channel (PCCH), the Common Traffic Channel (CTCH), the CCCH, the DCCH, and the Dedicated Traffic Channel (DTCH), respectively. A

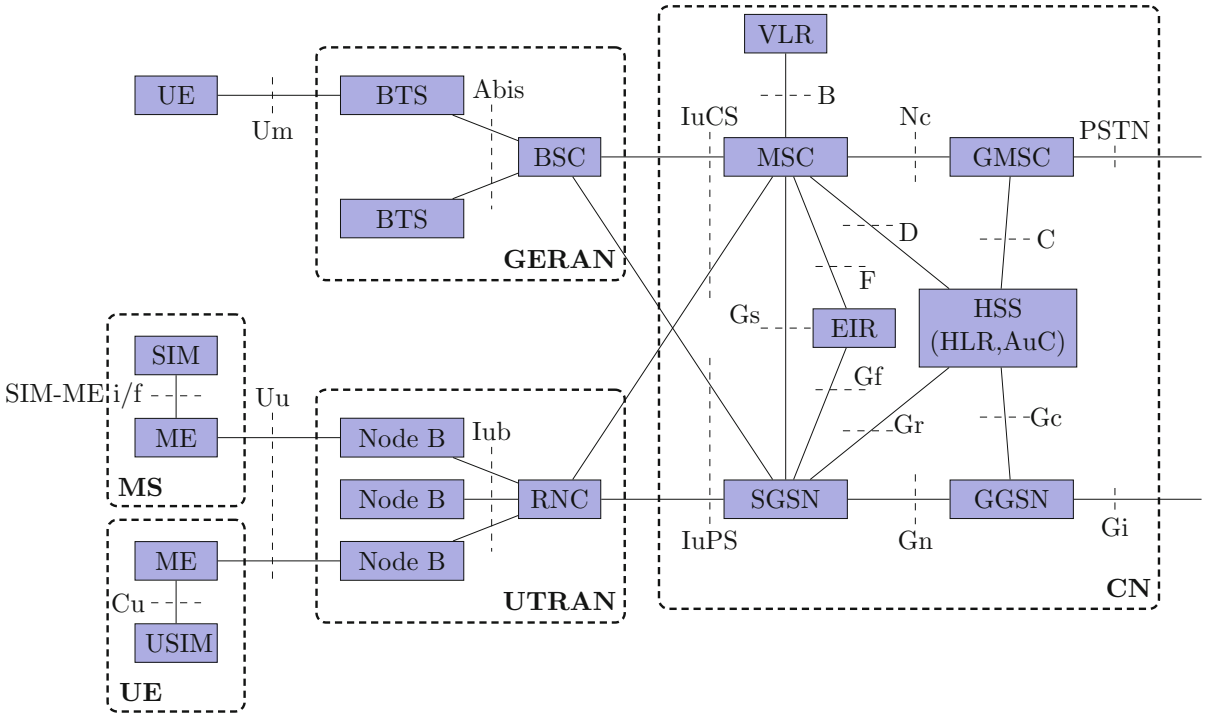


Fig. 6.1 The overall UMTS architecture and interfaces

Table 6.3 UMTS logical channels

	Abrv.	Name	Description
Broadcast	BCCH	Broadcast Control Channel	DL, for broadcasting system control information
Common	CCCH	Common Control Channel	DL, supports common procedures required to establish a dedicated link between the UE and the network
	CTCH	Common Traffic Channel	DL, a point-to-multipoint unidirectional channel for transfer of dedicated user information for all or a group of specified UEs
	PCCH	Paging Control Channel	DL, transfers paging information. Used when the network does not know the location cell of the UE or the UE is in sleep mode
Dedicated	DCCH	Dedicated Control Channel	UL/DL, a point-to-point dedicated channel for transmitting control information between a UE and the network
	DTCH	Dedicated Traffic Channel	UL/DL, a point-to-point dedicated channel for transmitting user traffic information between a UE and the network

categorized summary of them is provided in Table 6.3. To realize these logical channels, seven transport channels (TrCHs) are defined: the BCH, the PCH, the Forward Access Channel (FACH), the RACH, the Dedicated Channel (DCH), the HS-DSCH, and the E-DCH, which are described in Table 6.4. And finally, physical channels listed in Table 6.5 are defined and mapped onto the TrCHs. The channel mapping among the three layers is depicted in Fig. 6.2.

Independent of the type and direction of physical channel over which they are transmitted, all radio frames in UMTS share the same basic structure as shown in Fig. 6.3. Each frame lasts 10 ms, containing 15 slots of 2560 chips each (a chip is a basic time unit for UMTS radio transmission and has a duration of about 0.2604 µs). The radio format within each slot,

Table 6.4 UMTS transport channels

	Abrv.	Name	Description
Broadcast	BCH	Broadcast Channel	DL, used to broadcast system- and cell-specific information. Always transmitted over the entire cell and has a single transport format
Common	FACH	Forward Access Channel	DL, transmitted over the entire cell to carry data or information to the UEs that are registered on the system. There may be more than one FACH per cell as they may carry packet data
	PCH	Paging Channel	DL, always transmitted over the entire cell to carry messages that alert the UE to incoming calls, SMS messages, data sessions, or required maintenance such as re-registration
	RACH	Random Access Channel	UL, always received from the entire cell to carry requests for service from UEs trying to access the system
	HS-DSCH	High Speed Downlink Shared Channel	DL, shared by several UEs for high speed of transmission of bursty data packets and can be transmitted over the entire cell or over only part of the cell using, e.g., beamforming antennas
Dedicated	DCH	Dedicated Channel	UL/DL, used to transfer data to a particular UE. Each UE has its own DCH in each direction and can be transmitted over the entire cell or over only a part of the cell using, e.g., beamforming antennas
	E-DCH	Enhanced Dedicated Channel	UL, used to transfer user data with improved features such as higher data rates, reduced delays, and improved efficiency, for applications that require higher data rates in UL than in DL

however, is individually specified for every physical channel. For several types of physical channels, which are including the uplink DPCCH/DPDCH, the uplink HS-DPCCH, the uplink E-DPCCH/E-DPDCH, the HS-SCCH, the HS-PDSCH, and the E-AGCH, slots can also be clustered 3×3 into subframes of 2 ms each.

6.3.2 Spreading and Modulation

Spreading is a key technique of WCDMA that is used to increase the bandwidth of the signal and provide better security. The process involves two operations, namely channelization and scrambling. The channelization operation transforms each data symbol into multiple chips, thereby increasing the signal bandwidth. The number of chips per data symbol is referred to as the spreading factor (SF). During this process, data symbols on in-phase (I) and quadrature (Q) branches are multiplied independently with an orthogonal variable spreading factor (OVSF) code, where the I and Q denote real and imaginary parts, respectively. Afterward, in the scrambling step, a so-called scrambling code is applied to the spread signal, and the resultant signals on the I and Q branches are further multiplied by a complex-valued scrambling code.

6.3.2.1 Code Generation and Allocation

The generation of OVSF codes is based on a binary tree shown in Fig. 6.4. Each kth level of the tree generates a code set $\{c_{1,\mathrm{SF}}, c_{2,\mathrm{SF}} \ldots c_{\mathrm{SF},\mathrm{SF}}\}$, where $\mathrm{SF} = 2^{k-1}$. Both the selection of SF and the allocation of the codes onto different channels depend on the type of physical channel, which is specified in 3GPP TS 25.213 (2010).

There are two kinds of scrambling codes used in UMTS, namely the long and the short ones, respectively. The long scrambling codes are defined with a Gold-sequence (Zhang, 2011), which is a specific type of pseudo-random sequence with bounded small cross-correlations within a set. They can be efficiently generated with linear-feedback shift registers. The short codes, on the other hand, are defined from a sequence from the family of periodically extended S(2) codes. The allocation of scrambling codes to different physical channels is upon specification from higher layers.

Table 6.5 UMTS physical channels

		Abrev.	Full Name	Description
DL	Broadcast	CPICH	Common Pilot Channel	Used to carry a pre-defined bit sequence
		P-CCPCH	Primary Common Control Physical Channel	Used to carry the BCH transport channel
		SCH	Synchronization Channel	Used for cell search
		AICH	Acquisition Indicator Channel	Used to carry acquisition indicators
	Common	S-CCPCH	Secondary Common Control Physical Channel	Used to carry the FACH and PCH
		PICH	Page Indicator Channel	Used to carry the paging indicators
		HS-SCCH	Shared Control Channel for HS-DSCH	Used to carry downlink signaling related to HS-DSCH transmission
		HS-PDSCH	High Speed Physical Downlink Shared Channel	Used to carry the HS-DSCH
		E-AGCH	E-DCH Absolute Grant Channel	Used to carry the uplink E-DCH absolute grant
		MICH	MBMS Indicator Channel	Used to carry the MBMS notification indicators
	Dedicated	DPCH	Dedicated Physical Channel	Including both DPCCH and DPDCH like in the UL
		F-DPCH	Fractional Dedicated Physical Channel	Used to carry control information generated at Layer 1 (transmit power control commands)
		E-RGCH	E-DCH Relative Grant Channel	Used to carry the uplink E-DCH relative grants
		E-HICH	E-DCH Hybrid ARQ Indicator Channel	Used to carry the uplink E-DCH HARQ acknowledgement indicator
UL	Common	PRACH	Physical Random Access Channel	Used to carry the RACH
	Dedicated	DPCCH	Dedicated Physical Control Channel	Used to carry control information generated at Layer 1
		DPDCH	Dedicated Physical Data Channel	Used to carry the DCH transport channel
		HS-DPCCH	Dedicated Physical Control Channel (uplink) for HS-DSCH	Carries uplink feedback signaling related to downlink HS-DSCH transmission and to HS-SCCH orders
		E-DPCCH	E-DCH Dedicated Physical Control Channel	Used to carry the E-DCH transport channel
		E-DPDCH	E-DCH Dedicated Physical Data Channel	Used to carry the E-DCH transport channel

6.3.2.2 Uplink Spreading and Modulation

The overall scheme spreading of uplink dedicated channels is illustrated in Fig. 6.5: the DPCHs, the HS-DPCCH, and the E-DPCCH/E-DPDCHs are first separately spread, and then their complex sum is scrambled again as a whole data stream by the scrambling code $S_{\text{dpch},n}$. In the uplink, every dedicated physical channel can be scrambled with either a long code or a short code, while the PRACH is always scrambled with a long code. For the uplink there are in total 2^{24} long and 2^{24} short scrambling codes defined.

More specifically, the uplink DPCHs are spread as shown in Fig. 6.6. $c_{\text{d},1}$–$c_{\text{d},6}$ and $\beta_{\text{d},1}$–$\beta_{\text{d},6}$ are the channelization codes and gain factors of different DPDCHs, respectively. Similarly, c_{c} and β_{c} are the channelization code and gain factor for the DPCCHs.

The spreading of uplink HS-DPCCH is illustrated in Fig. 6.7, where the selection between mapping HS-DPCCH to the I branch and to the Q one is upon $N_{\text{max-dpdch}}$, the number of uplink DPDCHs. The setting of $N_{\text{max-dpdch}}$ is limited by the number of other uplink dedicated channels, as shown in Table 6.6.

The spreading of E-DPCCH/E-DPDCHs is shown in Fig. 6.8, where $c_{\text{ed},k}$, $\beta_{\text{ed},k}$, and $iq_{\text{ed},k} \in \{1, j\}$ are the channelization code, gain factor, and I/Q branch mapping value of the kth E-DPDCH, respectively. c_{ec}, β_{ec}, and iq_{ec} are similarly defined for the E-DPCCH.

In addition to the uplink dedicated channels, the common PRACH is also spread in a similar principle, as shown in Fig. 6.9. The detailed selection of channelization code, gain factor, and I/Q branch mapping value for each channel is specified in 3GPP TS 25.213 (2010).

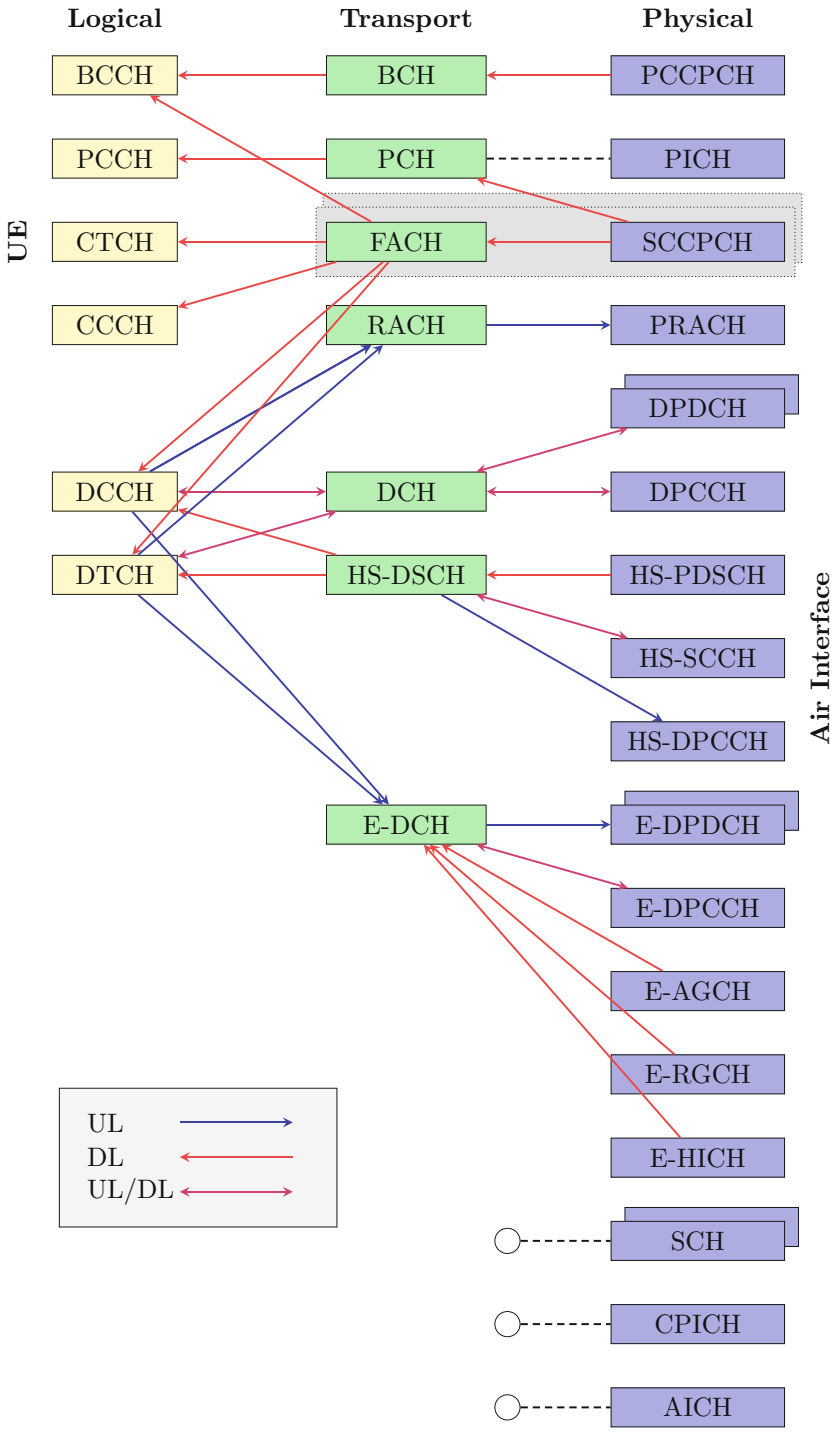


Fig. 6.2 UMTS channels in the UE perspective

6.3.2.3 Downlink Spreading

The spreading operation of all downlink physical channels except for SCH is illustrated in Fig. 6.10. For every individual physical channel, the I and Q branches are channelized by the same real-valued code $C_{\mathrm{ch,SF},m}$. Their complex-valued sum is then scrambled by a downlink scrambling code $S_{\mathrm{dl},n}$. For the downlink, a total of $\left(2^{18}-1\right)$ long scrambling codes can be generated, but only 8192 of them are actually used. These codes are divided into 512 sets, each consisting of one primary

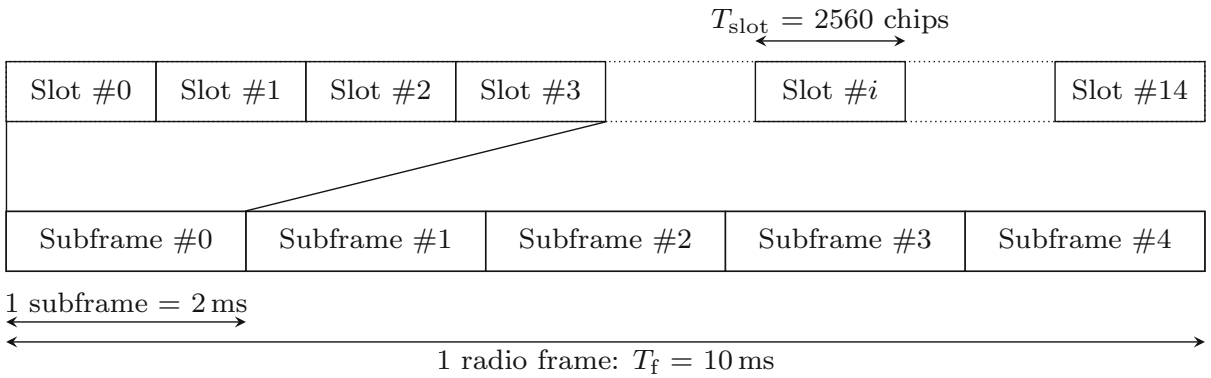


Fig. 6.3 Structure of UMTS radio frame

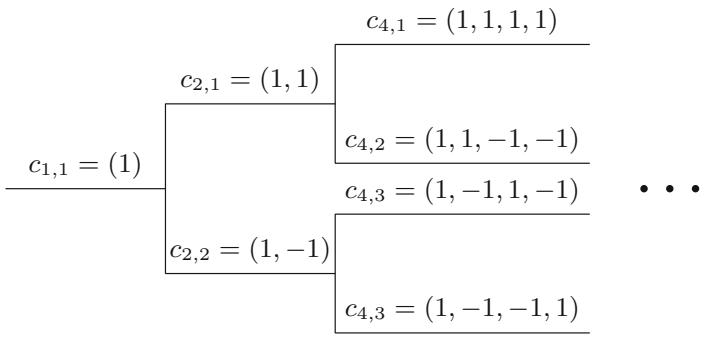


Fig. 6.4 Code tree for generation of OVSF codes

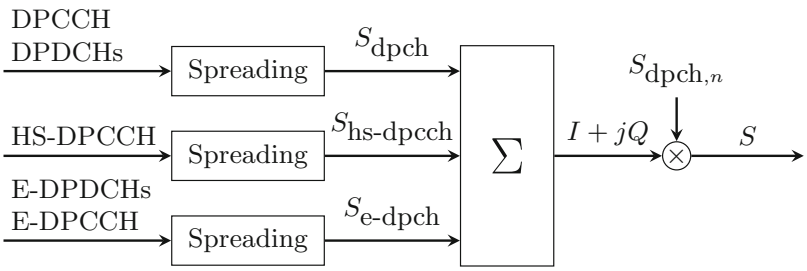


Fig. 6.5 Spreading for uplink dedicated channels

scrambling code and 15 secondary ones. The scrambled data streams of different DL physical channels, as illustrated in Fig. 6.11, are then combined together with the primary and secondary SCHs, where each individual channel is weighted by a gain factor.

6.3.2.4 Modulation

The modulation mechanism of UMTS physical channels can be generally described by Fig. 6.12. In the uplink, the input to modulation is the complex-valued chip sequence S from spreading; in the downlink, it is the chip sequence T from channel combining. For both cases, the modulating chip rate is 3.84 Mcps. By specifying the I/Q mapping, through which the chip sequence is split into I and Q branches, different modulation schemes can be realized, including BPSK, QPSK, pulse amplitude modulation (PAM), and QAM. More specifically, BPSK and 4-PAM can be used for uplink DPCHs; BPSK is used for PRACH; QPSK, 16-QAM, and 64-QAM can be used for HS-PDSCH and S-CCPCH; and QPSK is used for all other downlink channels except the SCH, which uses BPSK. The specifications of different I/Q mapping schemes are detailed in 3GPP TS 25.213 (2010). Both the I and Q branches are then modulated by a pulse shaping root-raised cosine (RRC) filter with roll-off coefficient $\alpha = 0.22$, of which the impulse response is

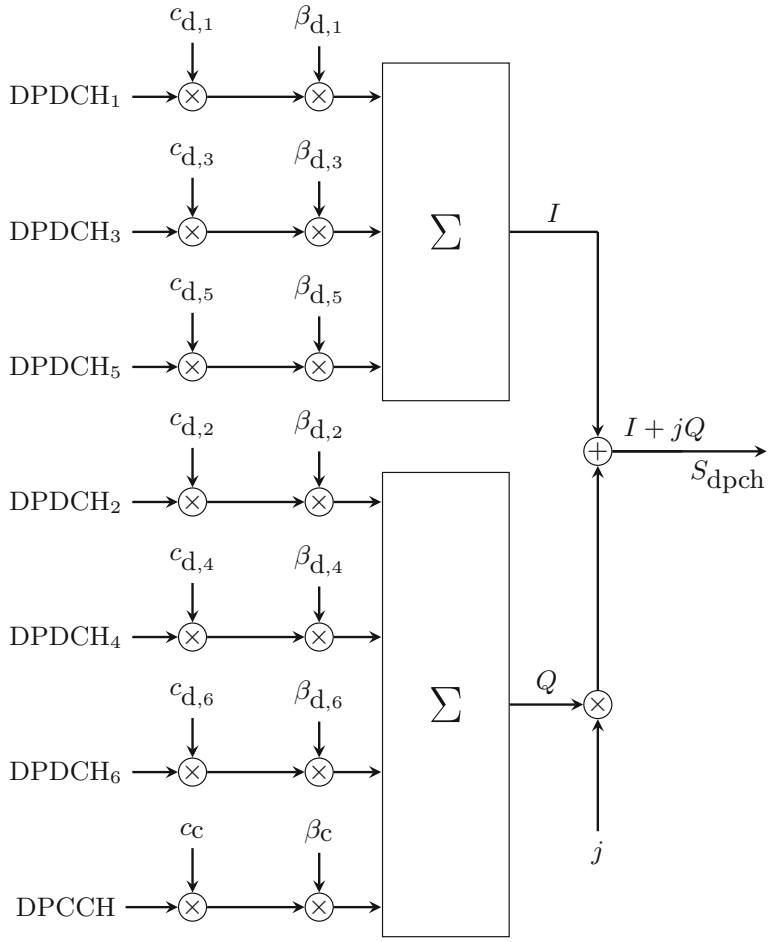


Fig. 6.6 Spreading for uplink DPCCH/DPDCHs

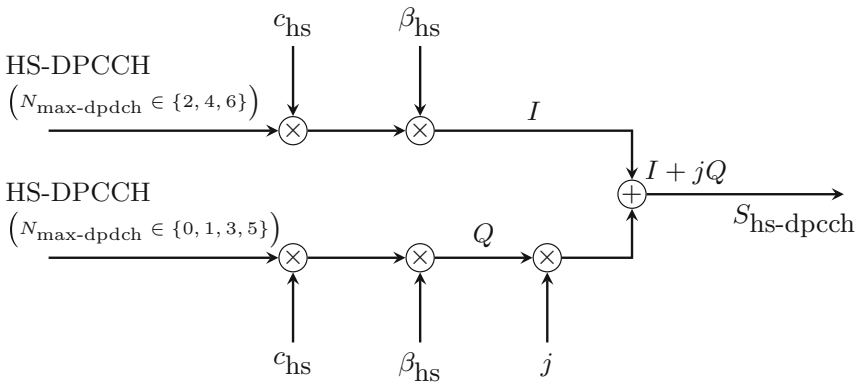


Fig. 6.7 Spreading for uplink HS-DPCCH

$$RC_0(t) = \frac{\sin\left[\pi \frac{t}{T_C}(1-\alpha)\right] + 4\alpha \frac{t}{T_C} \cos\left[\pi \frac{t}{T_C}(1+\alpha)\right]}{\pi \frac{den}{t} T_C \left[1 - \left(4\alpha \frac{t}{T_C}\right)^2\right]}. \tag{6.1}$$

The same RRC filter is used for both the uplink and the downlink.

Table 6.6 Maximum number of simultaneously configured UL DPCHs in UMTS

	DPDCH	HS-DPCCH	E-DPDCH	E-DPCCH
Case 1	6	1	–	–
Case 2	1	1	2	1
Case 3	–	1	4	1

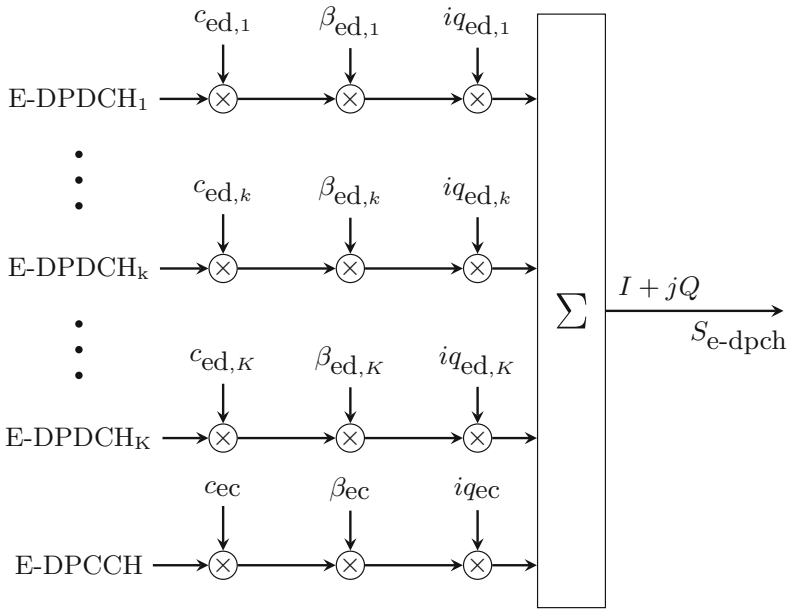


Fig. 6.8 Spreading for E-DPDCH/E-DPCCH

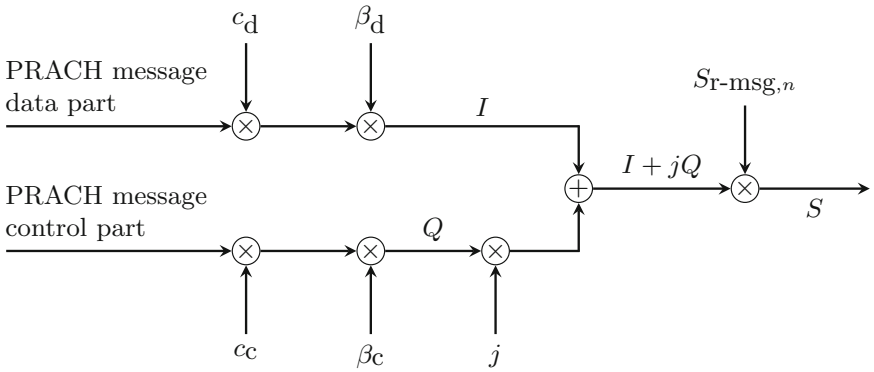


Fig. 6.9 Spreading for PRACH message part

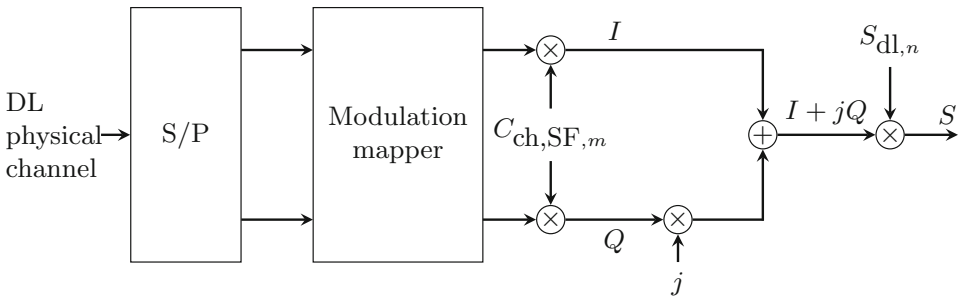


Fig. 6.10 Spreading for all downlink physical channels except SCH

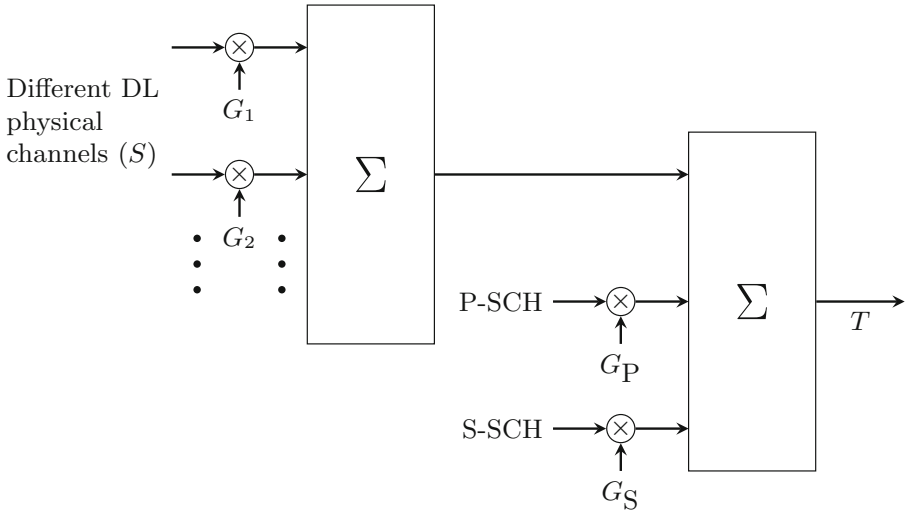


Fig. 6.11 Combining SCH with other downlink physical channels

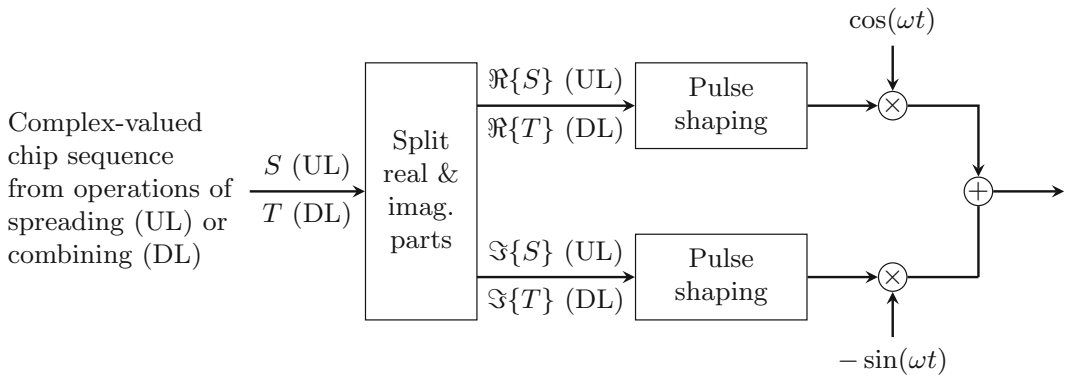


Fig. 6.12 UMTS modulation

6.3.3 Multiplexing, Channel Coding, and Interleaving

Data streams from/to MAC and higher layers are organized in the forms of transport block and transport block set. They shall be encoded/decoded to offer transport services over the radio transmission link. UMTS applies a combination of error detection, error correcting, rate matching, interleaving and TrCHs mapping onto/splitting from physical channels.

The transport channels DCH, RACH, BCH, FACH, and PCH share the same common channel coding and multiplexing scheme, which is shown in Figs. 6.13 and 6.14 for the uplink and downlink, respectively. Generally, CRC bits are first attacked to each transport block to provide error detection. The size of the CRC is 24, 16, 12, 8, or 0 bit(s), and the CRC size selection is signaled from higher layers for each TrCH. Afterward, all transport blocks in a TTI are serially concatenated. If the number of bits in a TTI is larger than the maximum size Z of a code block, a code block segmentation is performed after the concatenation of the transport blocks.

The code blocks are then sent to the channel encoder for FEC coding. UMTS uses two different coding schemes, namely the convolutional coding and the Turbo coding, respectively, for different transport channels (TrCHs), as listed in Table 6.7. Note that the maximal code block size Z also depends on the channel coding scheme. The implementation of the encoders is shown in Figs. 6.15 and 6.16, respectively. For detailed specification of the turbo code internal interleaver in Fig. 6.16, interested readers are referred to 3GPP TS 25.212 (2014).

From the output of channel coder, the multiplexing and coding chain becomes significantly different in UL and DL, as can be seen from Figs. 6.13 and 6.14. The reason for such asymmetry is twofold.

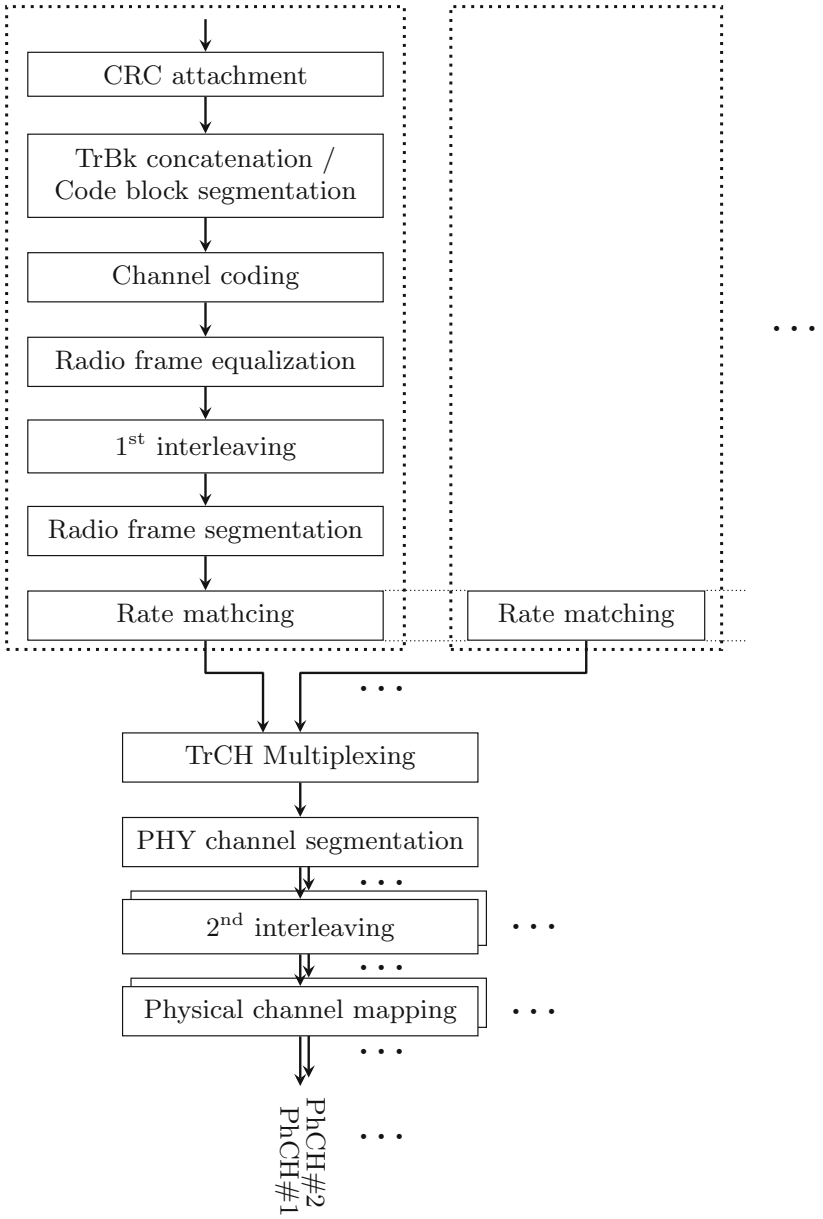


Fig. 6.13 UL multiplexing and coding chain of UMTS

First, the encoded data are organized and transmitted in radio frames, which are specified differently in the UL from that in the DL. First, the UL frames are generally smaller in size than the DL frames. The reason is that every UL frame comes from an individual UE, which has only limited transmission power and available resources. Therefore, to efficiently utilize the available resources and minimize interference, smaller frames are used in the UL. In contrast, with sufficient transmission power and radio resources, Node B is able to efficiently transmit larger frames in the DL. Moreover, in the UL, the bits available in each UL frame are individually determined for each TrCH and can be dynamically adjusted every TTI upon a set of parameters that specifies its format and transmission rate. The specification of these parameters is always restricted within a certain set called the Transport Format Set (TFS), and a specific combination of transport formats chosen from the TFS is known as the Transport Format Combination (TFC). In the DL, however, the frame size is not dependent on the TFC but given by the channelization code assigned by higher layers. Moreover, when the TTI is longer than 10 ms, an input data sequence must be segmented and mapped onto multiple radio frames with the same bit size. Thus, the bit size of the data sequence must always be an integer multiple of the frame size. In DL, it is accomplished by the rate matching, an operation of dynamically puncturing or repeating the data stream frame by frame, so as to adapt the TrCH data rate of the transport

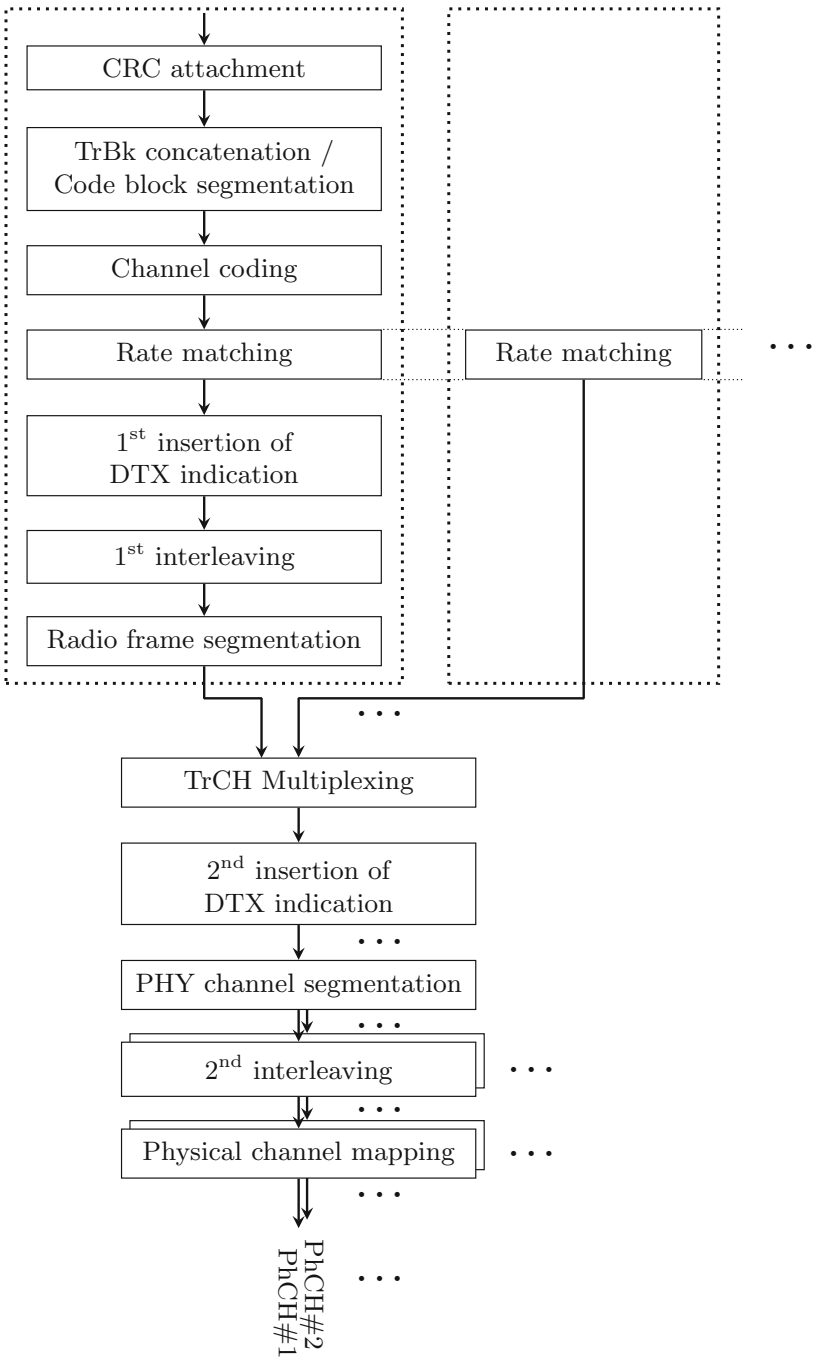


Fig. 6.14 DL multiplexing and coding chain of UMTS

channel to the instantaneously available channel capacity. Indeed, the flexibility of rate matching is limited, but yet sufficient to fulfill the segmentation requirement in the DL, since the frame size remains static over long terms. In the UL, however, to meet the segmentation requirement while the frame size is dynamically determined by the TFC, a more flexible approach is required to pad the coded bit sequence into an integer time of the frame size, which is known as the radio frame equalization.

Second, since the size of DL frames is generally large and remaining static, a waste of power and radio resource would occur when there is a silence period in the voice traffic. To mitigate such waste, DTX must be applied in the DL without having to break the long radio frames. In UMTS, the solution is to fill the silence period with DTX indication bits, which are detected by the transceiver and will not be transmitted. This procedure is carried out in the DL after rate matching, but unnecessary in the UL where the radio frame size is small and can be reconfigured every TTI.

Table 6.7 Usage and specification of channel coding schemes

Type of transport channel	Coding scheme	Max. code block size	Coding rate
BCH	Convolutional coding	504	1/2
PCH			
RACH			
DCH, FACH			1/3, 1/2
	Turbo coding	5114	1/3

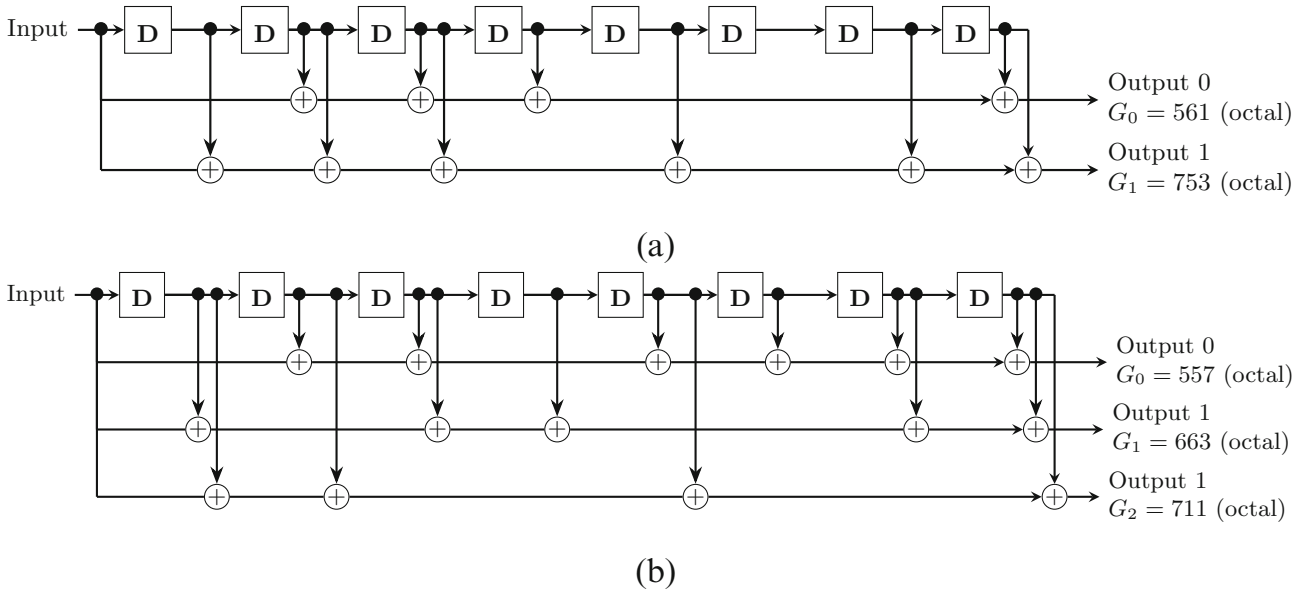


Fig. 6.15 Convolutional coders with rate of (**a**) 1/2 and (**b**) 1/3, respectively

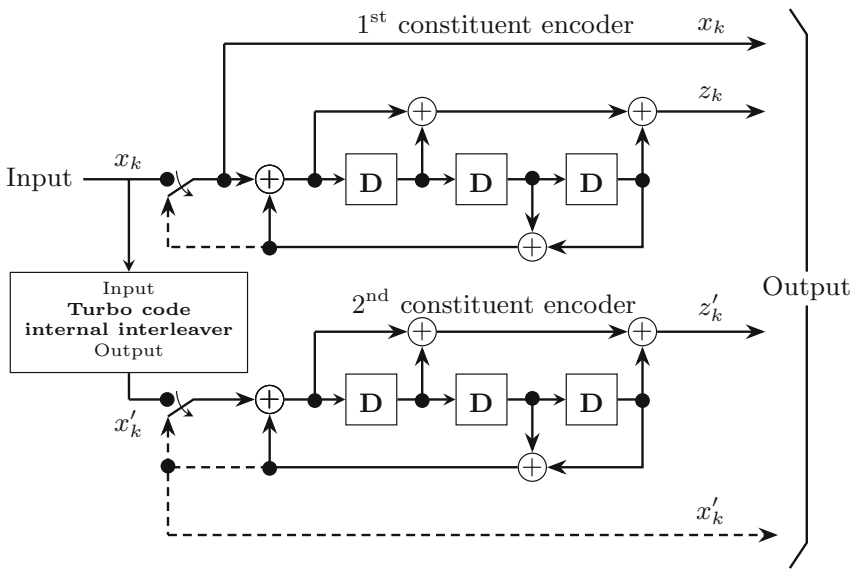


Fig. 6.16 Turbo coder with rate of 1/3 (dashed lines for trellis termination only)

After the radio frame equalization in UL and the first DTX indication insertion in DL, each TrCH is interleaved to spread the data across time and frequency domains, so as to provide robustness against burst errors caused by fading and other impairments in the radio channel. After this first interleaving, the radio frames are segmented.

After the segmentation (and rate matching in the UL), different TrCHs are multiplexed onto the same PHY channel(s). When more than one physical channel (PhCH) is used, the bits are then further segmented among the different PhCHs.

Before the segmented bits are finally mapped to the PhCH, a second interleaving is carried out. Since UMTS uses CDMA, this spreads the data across time, frequency, and code domains. Also due to the existence of this second interleaving, in the DL, the DTX indication bits must also be inserted for the second time, which is carried out between the TrCH multiplexing and the PHY channel segmentation.

6.3.4 Transport Format Detection

If the TFS of a TrCH contains more than one transport format, the transport format must be detected in the data flow. To achieve this, three possible methods are available, namely (i) TFC index (TFCI) based detection, which is applicable when the transport format combination is signaled using the TFCI field, (ii) explicit blind detection, which typically consists of detecting the transport format of the TrCH by use of channel decoding and CRC check, and (iii) guided detection, which is only applicable when there is at least one other TrCH that has the same TTI duration and transform formats and is under blind detection.

6.3.5 Compressed Mode

When a UE needs to perform a handover between cells with different frequencies, it must measure the signal quality of the target cell and report it to the network. This is achieved through inter-frequency measurement. However, implementing two RF transceivers on the UE to measure both the current and target cell signal quality would be impractical due to cost and potential interference issues.

To address this problem, UMTS uses compressed mode, which creates a small gap during which no transmission or reception takes place. This allows the UE to switch to the target cell and perform the signal quality measurement before returning to the current cell. The gap definition (starting position, gap length, number of gaps, etc.) must be agreed upon between the UE and network via RRC messages, such as Measurement Control and Physical Channel Reconfiguration. Three different methods can be used to achieve compressed mode operation: decreasing the spreading factor by 2:1, puncturing bits, and changing the higher layer scheduling to use fewer time slots for user traffic. The network decides which frames are compressed, and compressed frames can occur periodically or be requested on demand.

The compressed mode frame structure in the UL is standard, as illustrated in Fig. 6.17, but there are two different structures for the DL: Type A and Type B, which are used depending on the required transmission gap length and power control, as shown in Fig. 6.18. In Type A, the pilot field of the last slot in the transmission gap is transmitted, and transmission is turned off during the rest of the gap. In Type B, the TPC field of the first slot in the transmission gap and the pilot field of the last slot in the transmission gap are transmitted, and transmission is turned off during the rest of the gap.

6.3.6 Coding of HS-DSCH

Designed to provide high speed data services and requiring a higher data rate, the coding chain of HS-DSCH, as shown in Fig. 6.19, differs in several ways from that of the other TrCHs, which was discussed in Sect. 6.3.3. Firstly, it requires an extra bit scrambling before code block segmentation, which is to address the performance degradation from power level estimation for 16QAM when specific data sequences are transmitted on the HS-DSCH, e.g., all-zero or all-one sequences. Secondly, to efficiently counter the packet errors in high speed data transmission, HS-DSCH uses HARQ, of which the design is illustrated in Fig. 6.20. Thirdly, HS-DSCH does not use radio frame segmentation since the high data rates that it supports

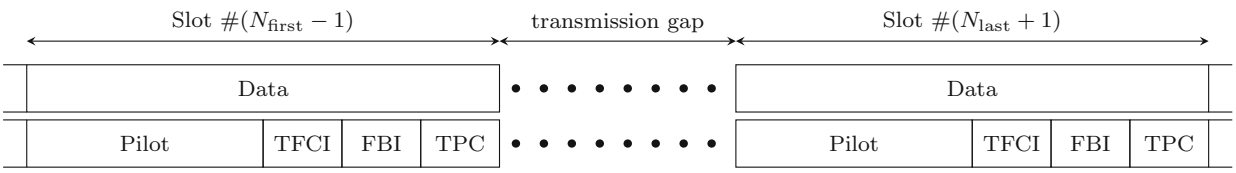


Fig. 6.17 Frame structure in UL compressed transmission

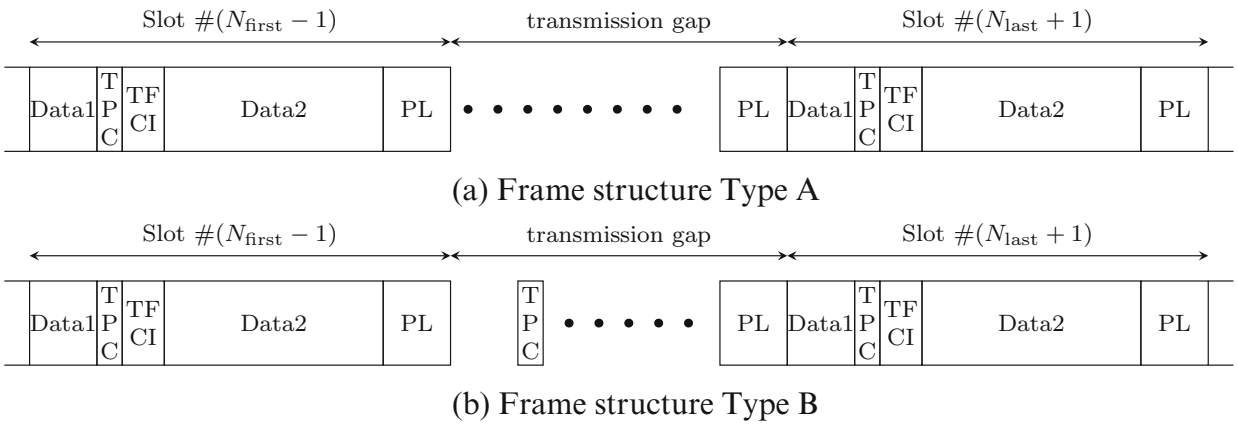


Fig. 6.18 Frame structure in DL compressed transmission. (**a**) Frame structure Type A. (**b**) Frame structure Type B

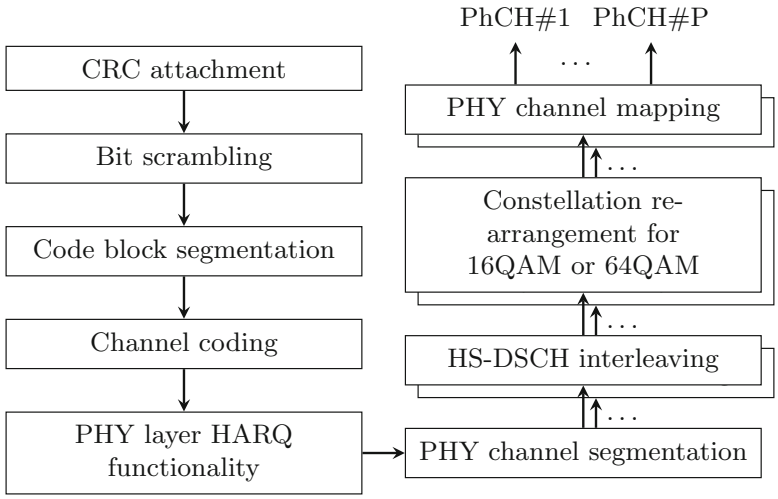


Fig. 6.19 HS-DSCH coding chain

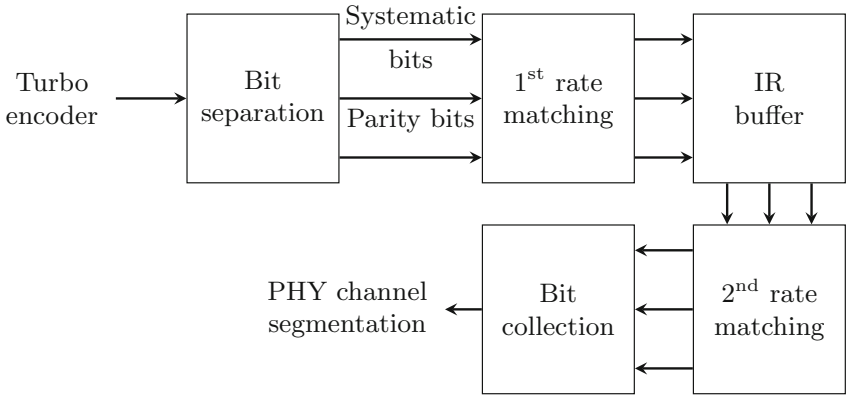


Fig. 6.20 HS-DSCH HARQ functionality

allow the entire transport block to fit into a single radio frame. Finally, HS-DSCH supports 16QAM and 64QAM, which are not used in other TrCHs. These higher-order modulations require more precise channel estimation and SNR than lower-order modulations and therefore requires a constellation rearrangement to improve the performance of the channel coding.

This difference makes the PHY design of HS-DSCH significantly different from the channels introduced in Sect. 6.3.3, making it complex and infeasible to multiplex them together. Instead, HS-DSCH is mapped to its own dedicated physical channels, to ensure high data rates and efficient transmission.

6.3.7 Coding of HS-SCCH

Upon the configuration of the UE, the HS-SCCH can be used in one of the three types.

More specifically, Type 1 is used when the UE is not configured in MIMO mode and the conditions for using Type 2 are not met. In this type, the HS-SCCH physical channel carries

- Seven channelization-code-set information bits that are always set to 1110000
- One modulation scheme information bit that is always set to 0 (for QPSK)
- Six transport block size information bit that are always set to 111101
- Three HARQ process information bits
- Three redundancy and constellation version bits
- One reserved bit for new data indicator
- Sixteen UE identity bits

Type 2 is used for the so-called HS-SCCH-less operation, which means that scheduling information for the HS-DSCH is transmitted on the HS-PDSCH instead of the HS-SCCH. This type is not used when the UE is configured in MIMO mode. For this type, there are at maximum three transmissions for an HS-DSCH transport block using CRC attachment method 2, without an associated HS-SCCH. Meanwhile, the HS-SCCH Type 2 physical channel is used to transmit the following information during the second and third transmissions:

- Channelization-code-set information (7 bits)
- One modulation scheme information (1 bit)
- Special information type (6 bits)
- Special information (7 bits)
- UE identity (16 bits)

Type 3 is used when the UE is configured in MIMO mode. For this type, if one transport block is transmitted on the associated HS-PDSCH(s) or an HS-SCCH order is transmitted, the HS-SCCH physical channel transmits:

- Channelization-code-set information (7 bits), always set to 1110000
- Modulation scheme and number of transport blocks information (3 bit), always set to 000
- Precoding weight information (2 bits), always set to 00
- Transport block size information (6 bits), always set to 111101
- HARQ process information (4 bits)
- Redundancy and constellation version (2 bits)
- UE identity (16 bits)

If two transport blocks are transmitted on the associated HS-PDSCHs, the HS-SCCH Type 3 physical channel transmits

- Channelization-code-set information (7 bits)
- Modulation scheme and number of transport blocks information (3 bits)
- Precoding weight information for the primary transport block (2 bits)
- Transport block size information for the primary and secondary transport blocks, respectively (each 6 bits)
- HARQ process information (4 bits)
- Redundancy and constellation version for the primary and secondary transport blocks, respectively (each 2 bits)
- UE identity (16 bits)

In all HS-SCCH types, the information bits are first clustered into two groups, and two parallel channel coding paths are implemented: path 1 encodes all bits in group 1 and applies UE specific masking and path 2 encodes all bits in both groups while adding UE specific CRC attachment. The two paths are matched in rate and combined before being mapped onto the HS-SCCH physical channel. The detailed specification of the coding chain is configured upon the HS-SCCH type, as described in 3GPP TS 25.212 (2014).

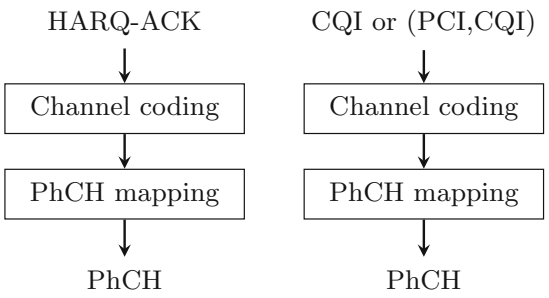


Fig. 6.21 Coding for HS-DPCCH

6.3.8 Coding of HS-DPCCH

The coding for HS-DPCCH, as illustrated in Fig. 6.21, has two parallel chains, both consisting of a channel coding stage and a PhCH mapping. The first chain processes the acknowledge message of HARQ, while the second is specified upon the configuration of UE. When the UE is not working in MIMO mode, it only encodes a 5-bit CQI; otherwise it must also encode the precoding control indication (PCI). The PCI can be either 5 bits long when the CQI is of Type A, or 2 bits long for Type B CQI, as specified in 3GPP TS 25.214 (2010).

6.3.9 Physical Layer Procedures

The physical layer procedures of FDD UMTS are specified in 3GPP TS 25.214 (2010), mainly including the synchronization, the power control, and the random access.

6.3.9.1 Synchronization

In UMTS, the synchronization begins with the cell search, when the UE searches for a cell and determines the DL scrambling code and frame synchronization of that cell. The cell search is typically carried out in three steps:

1. The slot synchronization, where the UE uses the SCH's primary synchronization code, which is common to all cells, to acquire slot synchronization to a cell. This is typically done with a single matched filter that is matched to the primary synchronization code.
2. The frame synchronization and code group identification, where the UE uses the SCH's secondary synchronization code to find frame synchronization and identify the code group of the cell found in the first step. This is typically done by correlating the received signal with all possible secondary synchronization code sequences and identifying the maximum correlation value.
3. Having identified, the code group, the UE determines the exact primary scrambling code used by the found cell, which is typically done through a symbol-by-symbol correlation over the CPICH with all codes within the identified code group.

After cell search, the frame timing of all common physical channels can be determined. It is worth noting that when multiple S-CCPCHs on different radio links can be soft combined, timing information will be provided by higher layers so that the UE is allowed to determine the time interval of combining that applies to each S-CCPCH, which is known as the *L1 combining period*.

For the dedicated channels, it takes synchronization primitives, in both UL and DL, to indicate the synchronization status of radio links. In the DL, the status is determined based on the UE-estimated quality of the Transmit Power Control (TPC) fields of the F-DPCH frame (when not in UL-DTX mode), the UE-estimated quality of the DPCCH (when in UL-DTX mode), and the CRC checking result in the DPDCH. In the UL, the indicating criteria is not specified by 3GPP but shall be generally similar to those of the DL. Upon these primitives, each radio link set is considered in Node B to be in one among three different states: initial state, in-sync state, and out-of-sync state. The transitions between different states are illustrated in Fig. 6.22.

Starting from the initial state, two procedures are defined in order to obtain physical layer synchronization of dedicated channels between UE and UTRAN, namely the synchronization procedures A and B, respectively.

The synchronization procedure A is used when at least one DL dedicated physical channel and one UL dedicated physical channel are to be set up on a frequency and none of the radio links after the establishment/reconfiguration existed prior to the establishment/reconfiguration. In this approach, Node B first sets up all radio link sets to be set up for the UE in the initial

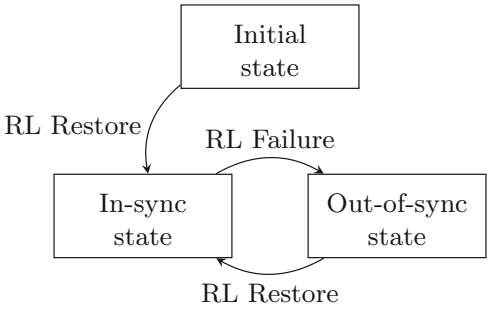


Fig. 6.22 Node B radio link set states and transitions

state; then UTRAN initiates downlink transmission so that UE can synchronize with the DL channels. Then UE starts an uplink transmission immediately or after a post-verification period, so that UTRAN can establish the uplink synchronization.

The synchronization procedure B is used, in contrast, when one or several radio links are added to the active set and at least one of the radio links prior to the establishment/reconfiguration still exists after the establishment/reconfiguration. In this case, new radio link sets are set up to be in the initial state, and the existing radio link set which gets new radio link(s) added shall be considered to remain in its previous state. UTRAN starts DL transmission and simultaneously establishes UL synchronization for each new radio link, while the UE synchronizes with each new radio link's DL channels.

The transitions between in-sync and out-of-sync states are triggered by radio link failure/restore procedures, upon the result of radio link monitoring.

Additionally, during a connection, the UE may autonomously adjust its DPDCH/DPCCH transmission time instant. When it does so, it shall also simultaneously adjust the HS-DPCCH, E-DPCCH, and E-DPDCH transmission time instant by the same amount, so that DPCCH/DPDCH remain in time alignment with E-DPCCH/E-DPDCH and that their relative timing with HS-DPCCH remains constant.

6.3.9.2 Power Control

The UTRA FDD physical layer uses multiple approaches for power control.

When a UE is accessing the network, an open-loop power control is applied for setting the initial transmit powers in UL and DL. Lacking accuracy, it tolerates up to ± 9 dB error in normal conditions or ± 12 dB in extreme conditions.

After accessing the network, the UL transmit power of UE is initialized by higher layers and then adjusted based on the feedback from Node B, so as to keep the received uplink SIR at a given target level. This power adjustment is the inner loop power control, a.k.a. the fast closed-loop power control. The target SIR level, on the other hand, is determined by the outer loop power control so as the maintain the communication quality.

The inner loop control is executed upon TPC commands generated by Node B, one command per slot, with a minimum adjustment step size of 1 dB. Larger step sizes can be used, and smaller step size can be "emulated" by applying the minimum step size once per multiple slots. Similarly, in the DL, the network determines the transmit power for DPCCH, DPDCH, and F-DPCH. The UE generates TPC commands to control the network transmit power and send them in the TPC field of the uplink DPCCH. The power control loop adjusts the power of the DPCCH and DPDCHs with the same amount, so as to keep the relative power difference between them, which is determined by the network.

For HS-SCCH, HS-PDSCH, E-AGCH, E-HICH, and E-RGCH, the DL power control is under control of Node B. For AICH, PICH, S-CCPCH, and MICH, the UE is informed about the relative transmit power compared to the primary CPICH transmit power by the higher layers.

Special situations such as compressed mode and handover can create additional issues for the power control, which are specified with details in 3GPP TS 25.214 (2010).

6.3.9.3 Random Access

In UTRA, before the physical random access procedure can be initiated, the PHY layer must receive some essential information from the RRC layer, including (i) the preamble scrambling code, (ii) the message length in time, either 10 or 20 ms, (iii) the AICH transmission timing parameter, (iv) the set of available signatures and the set of available RACH sub-channels for each access service class (ASC), (v) the power-ramping factor, (vi) the maximum preamble retransmission times (vii) the initial preamble power, (viii) the offset between the power of the last transmitted preamble and the control part of the random access message, and (ix) the set of Transport Format parameters.

With the information received/updated at the PHY layer, the MAC sublayer can request to initiate a physical random access procedure, by which it sends to the PHY layer (i) the Transport Format to be used for the PRACH message part, (ii) the ASC of the PRACH transmission, and (iii) the data to be transmitted (Transport Block Set).

The physical random access procedure is performed by the UE as follows:

1. Derive the available uplink access slots for the set of available RACH sub-channels within the given ASC, and randomly select one among them.
2. Select a random signature from the set of available signatures within the given ASC.
3. Reset the preamble retransmission counter, and select the preamble transmission power.
4. Transmit a preamble using the selected uplink access slot, signature, and preamble transmission power.
5. In case no ACK or NACK is detected on AICH in the corresponding DL access slot, reselect a random signature, increase the preamble transmission power, and retransmit the preamble in the next available access slot. The retransmission repeats until detecting an acquisition indicator or exceeding the maximum retransmission times.
6. In case of contention (AICH NACK), terminate the procedure.
7. Upon detection of ACK on AICH, transmit the RACH message part to accomplish the procedure.

6.4 Radio Interface Protocols

To facilitate efficient and reliable communication between UE and UTRAN, UMTS has introduced a new protocol stack, which consists of several sublayers, as shown in Fig. 6.23. Focusing on the lower layers of the communication system, it contains six (sub)layers over the three bottom layers. More specifically:

1. The PHY layer (Layer 1) handles the transmission and reception of raw bitstream over the physical medium. It includes tasks such as modulation, coding, and radio frequency transmission. This layer is responsible for converting digital data into physical signals for transmission and vice versa.
2. The MAC sublayer of Layer 2 manages access to the shared radio channel. It handles tasks such as channel allocation, multiplexing, and error control. The MAC sublayer ensures efficient and fair utilization of the available radio resources.
3. The RLC sublayer of Layer 2 provides reliable transmission of data segments between the transmitting and receiving entities. It handles segmentation, reassembly, error detection, and retransmission of data to ensure reliable data delivery.

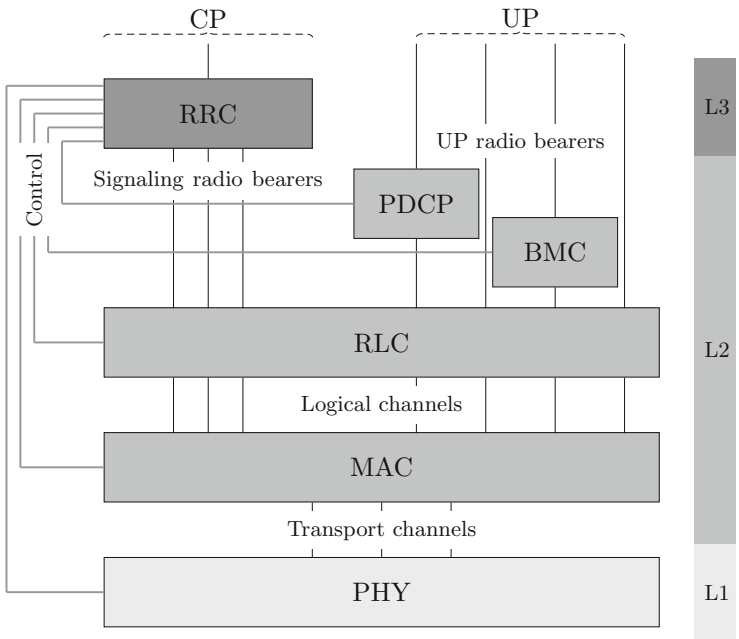


Fig. 6.23 UMTS radio interface protocol stack

4. The Packet Data Convergence Protocol (PDCP) sublayer of Layer 2 is responsible for functions such as header compression, encryption, and integrity protection for user data. It ensures the efficient and secure transfer of data packets between the transmitting and receiving entities.
5. The broadcast/multicast control (BMC) sublayer of Layer 2 manages the distribution of broadcast and multicast information to multiple users. It handles the efficient delivery of data to multiple recipients and supports the management of broadcast channels.
6. The RRC sublayer of Layer 3 manages the control signaling between the mobile device and the network. It handles tasks such as radio resource allocation, connection establishment, and mobility management. It ensures the necessary signaling for controlling the radio access network.

While the UMTS protocol stack may not perfectly align with the classical Open Systems Interconnection (OSI) model (International Organization for Standardization, 1994), we can draw some similarities. The UMTS Layer 1 can be compared to the PHY layer in the OSI model, as they both handle the physical transmission of data. The sublayers in UMTS Layer 2 collectively provide functionalities similar to the data link layer and parts of the transport layer in the OSI model. The RRC sublayer in UMTS, since it performs functions related to connection management and control, is somewhat resembling the network layer in the OSI model.

In addition to the vertical protocol stack, UMTS also introduces the concept of control/user splitting. The control plane (CP) carries signaling and control information, while the user plane (UP) carries user data transmission. This splitting allows for the independent management and optimization of signaling/control functions and user data transmission, enhancing the efficiency and performance of the network.

6.4.1 The MAC Sublayer

6.4.1.1 MAC Sublayer Architecture

Logically, the MAC sublayer is divided into three entities, each being responsible for a type of transport channel(s): (i) the MAC-b for the broadcast channel, (ii) the MAC-c/sh for the common and shared channels, and (iii) the MAC-d for the dedicated channels. The logical architecture can be briefly illustrated by Fig. 6.24.

6.4.1.2 MAC Functions and Services

The functionalities provided by the MAC sublayer include:

1. Mapping between logical channels and transport channels
2. Selection of appropriate transport format for each transport channel depending on instantaneous source rate

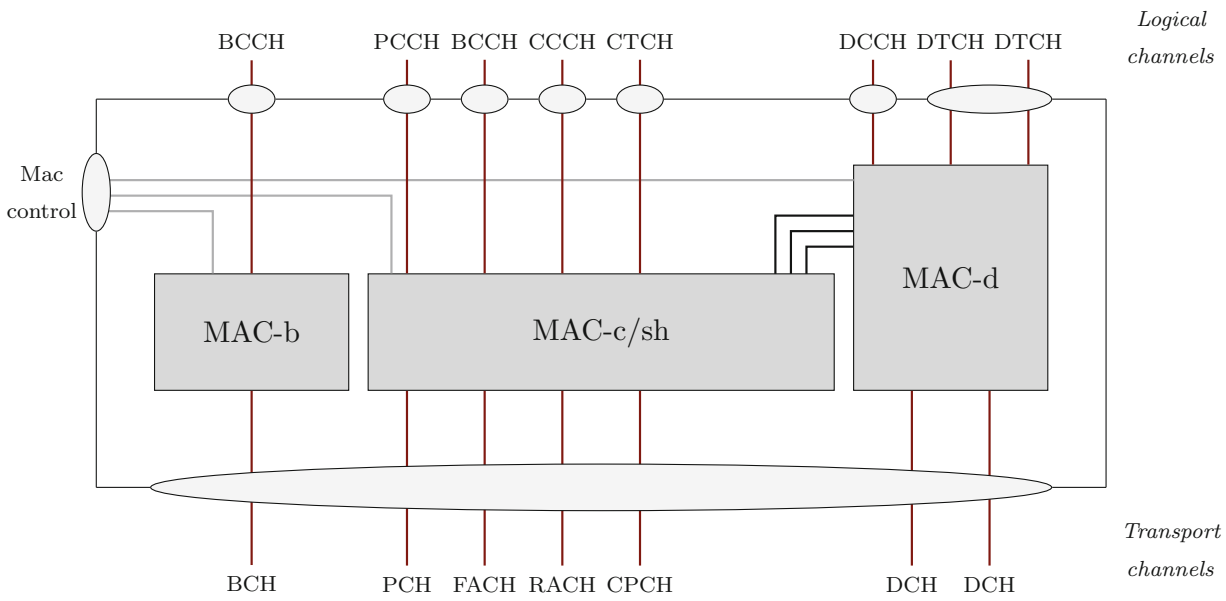


Fig. 6.24 The UMTS MAC sublayer architecture

3. Priority handling between data flows of one UE
4. Priority handling between UEs by means of dynamic scheduling
5. Identification of UEs on common transport channels
6. Identification of Multimedia broadcast/multicast service (MBMS) on common transport channels
7. Multiplexing/demultiplexing of upper layer Protocol Data Units (PDUs) into/from transport blocks delivered to/from the physical layer on common transport channels
8. Multiplexing/demultiplexing of upper layer PDUs into/from transport block sets delivered to/from the physical layer on dedicated transport channels
9. Segmentation and reassembly of upper layer PDUs—traffic volume measurement
10. transport channel type switching
11. Ciphering for transparent mode RLC
12. ASC selection for RACH transmission
13. Control of HS-DSCH transmission and reception including support of HARQ
14. HS-DSCH provided bit rate measurement
15. Control of E-DCH transmission and reception including support of HARQ
16. Generation of uplink scheduling information to assist with E-DCH resource allocation
17. E-DCH provided bit rate measurement

6.4.2 The RLC Sublayer

6.4.2.1 RLC Modes and Overall Model

On the RLC sublayer, each instance is configured by the RRC sublayer to operate in one of the three modes: the transparent mode (Tr), the unacknowledged mode (UM), or the acknowledged mode (AM). More specifically:

- In Tr, the RLC entity is implemented with the lowest complexity and minimal functionality. It adds no protocol overhead, such as sequence numbers or acknowledgments, to higher layer data. The higher layer data can be transmitted in streaming mode, i.e., without being segmented. Nevertheless, some limited segmentation and reassembly can still be accomplished in special cases. Tr is suitable for applications that require minimal RLC processing, such as voice services.
- UM introduces sequence numbers to observe the integrity of higher layer PDUs, so as to provide a higher level of reliability than Tr does. However, it does not include any acknowledgment mechanism for the received data units, so the data delivery is not guaranteed either. It is suitable for applications with moderate reliability requirements and real-time constraints, such as streaming multimedia.
- AM offers the highest level of reliability among the three RLC modes. It includes sequence numbers, acknowledgments, and retransmissions to guarantee the delivery of data units. AM provides error detection, retransmission mechanisms, and flow control, ensuring reliable data transmission. It is commonly used for applications that require reliable and error-free data transfer, such as web browsing and file downloads.

RLC entities are defined unidirectional for Tr and UM, but bidirectional for AM. Therewith, the overall model of the RLC sublayer is defined as shown in Fig. 6.25.

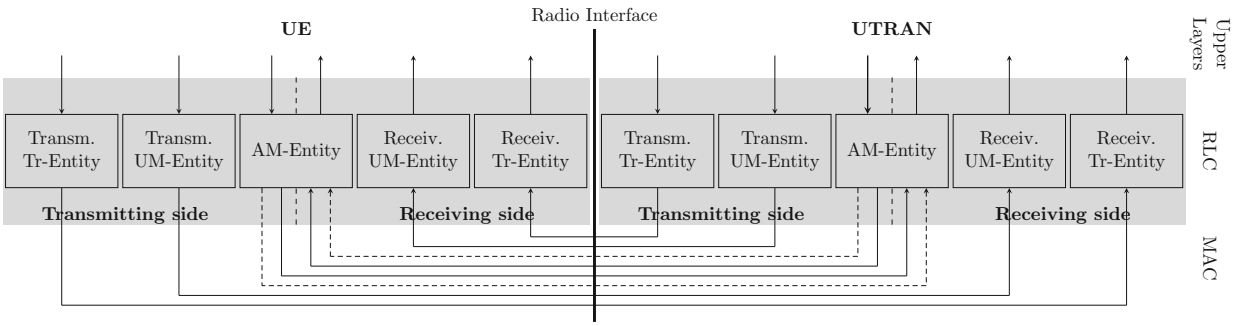


Fig. 6.25 The UMTS RLC sublayer model

6.4.2.2 RLC Functions

The functionalities provided by the RLC sublayer include:

1. Segmentation and reassembly
2. Concatenation
3. Padding
4. Transfer of user data
5. Error correction
6. In-sequence delivery of upper layer PDUs
7. Duplicate detection
8. Flow control
9. Sequence number check
10. Protocol error detection and recovery
11. Ciphering
12. service data unit (SDU) discard
13. Out of sequence SDU delivery
14. Duplicate avoidance and reordering

6.4.3 The PDCP Sublayer

6.4.3.1 PDCP Sublayer Architecture

The PDCP sublayer, defined for the packet switched (PS) domain only, is structured within the radio interface protocol architecture as shown in Fig. 6.26. Every PS domain radio access bearer (RAB) is associated with one radio bearer (RB), which in turn is associated with one PDCP entity. Each PDCP entity is associated with one bidirectional RLC entity or two directional ones. Depending on the configuration, every PDCP entity uses zero, one, or several different header compression (HC) protocols—and of each HC protocol at most one instance. Several PDCP entities may be defined for a UE with each using the same or a different set of header compression protocols. The PDCP sublayer is configured by upper layer through the PDCP-Control SAP (C-SAP).

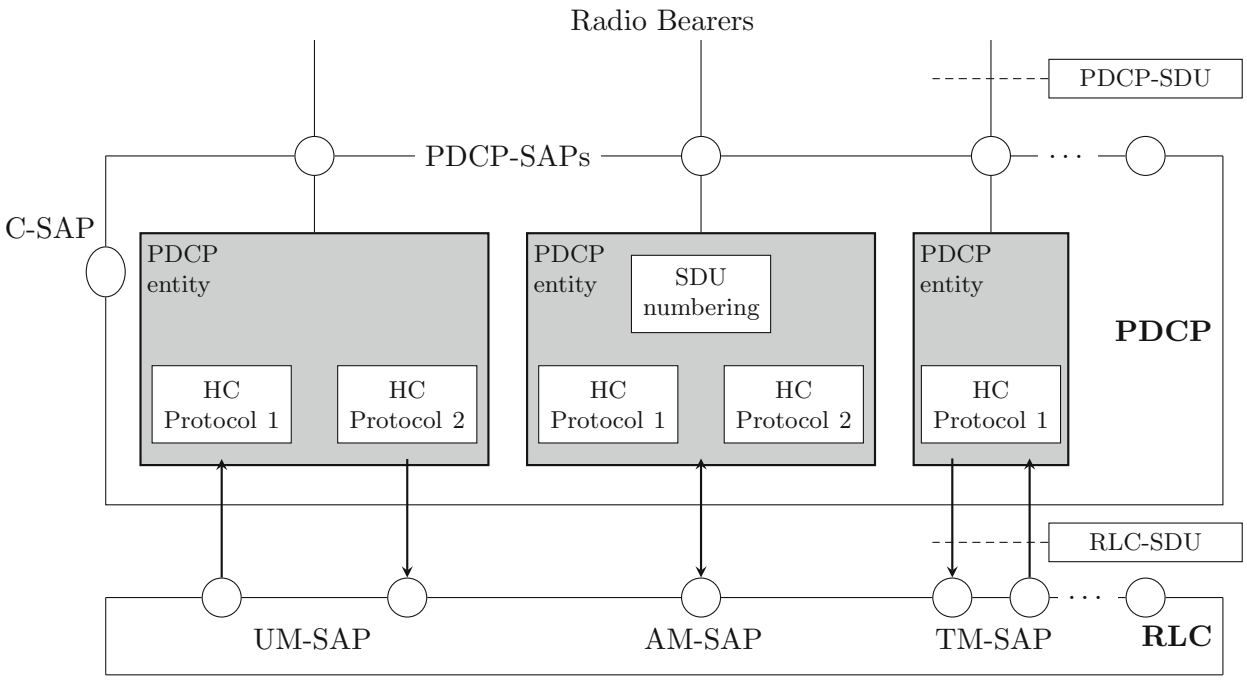


Fig. 6.26 The UMTS PDCP architecture

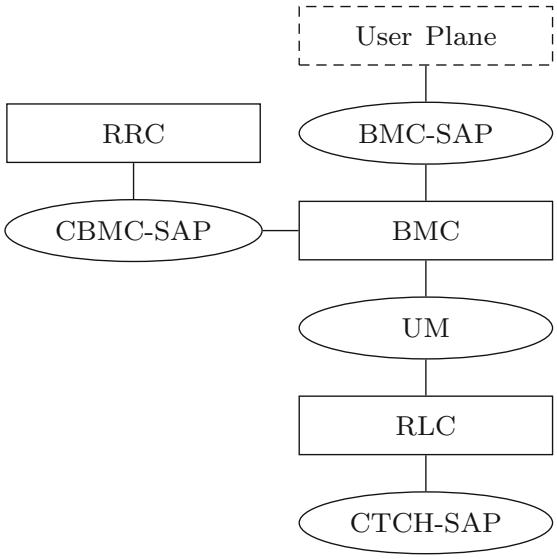


Fig. 6.27 The UMTS BMC model

6.4.3.2 PDCP Functions

The PDCP sublayer provides the following functionalities:

1. Compression/decompression of redundant protocol control information at the transmitting/receiving entities, respectively
2. Transfer of user data
3. Support for lossless Serving Radio Network Subsystem (SRNS) relocation

6.4.4 The BMC Sublayer

6.4.4.1 BMC Sublayer Model

The model of the BMC sublayer is illustrated by Fig. 6.27. At the UTRAN side, the BMC sublayer shall consist of one protocol entity per cell, each requiring a single CTCH that is provided by the MAC sublayer through the RLC sublayer. The BMC requests the UM service of the RLC. There is assumed to be a function in the RNC above BMC that resolves the geographical area information of the cell broadcast (CB) message received from the Cell Broadcast Center (CBC). A BMC protocol entity serves only those messages at BMC-Service Access Point (SAP) that are to be broadcast into a specified cell.

6.4.4.2 BMC Functions

The functions provided by the BMC sublayer include:

1. Storage of CB message
2. Traffic volume monitoring and radio resource request for cell broadcast service (CBS)
3. Scheduling of BMC messages
4. Transmission of BMC messages to UE
5. Delivery of CB messages to the upper layer

6.4.5 The RRC Sublayer

6.4.5.1 RRC Sublayer Model

The RRC sublayer is logically defined as shown in Fig. 6.28. Routing of higher layer messages to different MM/connection management (CM) entities (UE side) or different core network domains (UTRAN side) is handled by the Routing Function Entity (RFE). Within the RRC sublayer are four different entities, namely Broadcast Control Functional Entity (BCFE), Paging and Notification Control Functional Entity (PNFE), Dedicated Control Functional Entity (DCFE), and Shared Control Functional Entity (SCFE). The BCFE is used to handle broadcast functions and deliver the RRC services, which are required

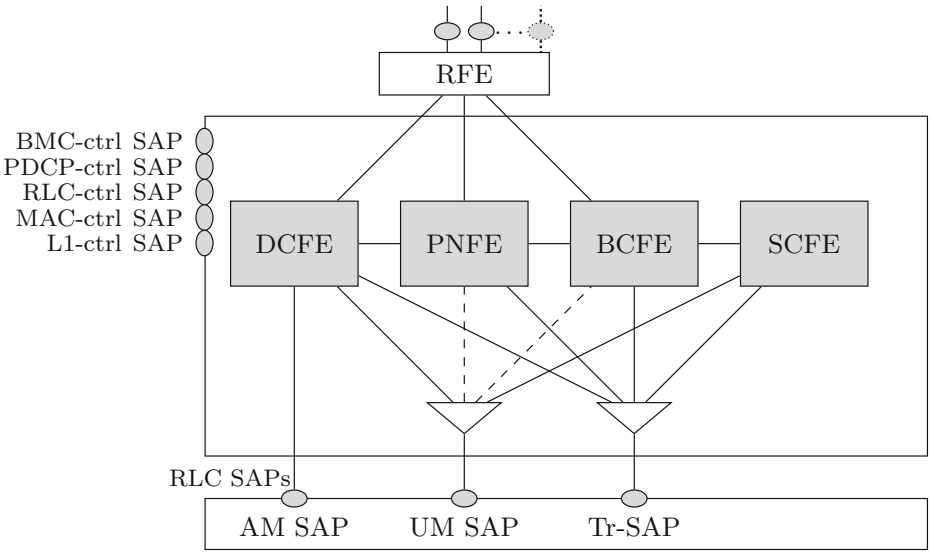


Fig. 6.28 The UMTS RRC sublayer architecture

at the General Control SAP (GC-SAP). It can use the lower layer services provided by the Tr-SAP and UM-SAP. The PNFE is used to control the paging of UEs that do not have an RRC connection and deliver the RRC services that are required at the Notification SAP (Nt-SAP). It can use the lower layer services provided by the Tr-SAP and UM-SAP. The DCFE handles all functions specific to one UE and delivers the RRC services that are required at the Dedicated Control SAP (DC-SAP). Depending on the message to be sent and on the current UE service state, it can use lower layer services of UM/AM-SAP and Tr-SAP.

6.4.5.2 RRC States

In UMTS, after being switched on and registered to a PLMN, a UE can be in one of five different RRC states, namely the IDLE state, the CELL_FACH state, the CELL_DCH state, the CELL_PCH state, and the URA_PCH state, respectively. Especially, the latter four are cumulatively called the RRC connected mode. The transitions between these states are illustrated in Fig. 6.29. More specifically:

- In the IDLE state, the UE is not actively involved in any communication with the network. It is able to monitor the system information over the BCH. When a connection request is initiated or system information needs to be acquired, the device establishes an RRC connection and transitions to the CELL_FACH or CELL_DCH state, depending on the data rate it requires.
- In the CELL_DCH state, the UE is actively engaged in data transfer or signaling with the network. It is assigned with a DPCH and known by its serving RNC on a cell or active set level. In this state, the UE performs measurements and sends measurement reports according to the measurement control information received from RNC. When there is no ongoing data transfer or signaling and the device transitions to a lower power mode, it moves to the CELL_FACH state. If the device needs to conserve power during inactivity, it transitions to the CELL_PCH or URA_PCH state. When the device completes its active communication or when there is a loss of coverage, it releases the RRC connection and transitions to the IDLE state.
- In the CELL_FACH state, the UE is connected to the network, but not assigned with any DCH. Instead, only RACHs and FACHs are used for signaling messages and small amounts of UP data. The UE is also capable of listening to the BCH to receive system information. In this state UE performs cell reselection and cell update and is therefore known by the RNC on a cell level. If higher data rates are required for data transfer or signaling purposes, the device transitions to the CELL_DCH state. If the device needs to conserve power during inactivity, it transitions to the CELL_PCH or URA_PCH state. When the data transmission requirements decrease and there is no ongoing communication, it releases the RRC connection and transitions to the IDLE state to conserve power and release network resources.
- In the CELL_PCH state, the UE enters a low-power mode to conserve battery during periods of inactivity. It executes cell reselection and cell update and is therefore still known by the network on the cell level, but only reachable via the

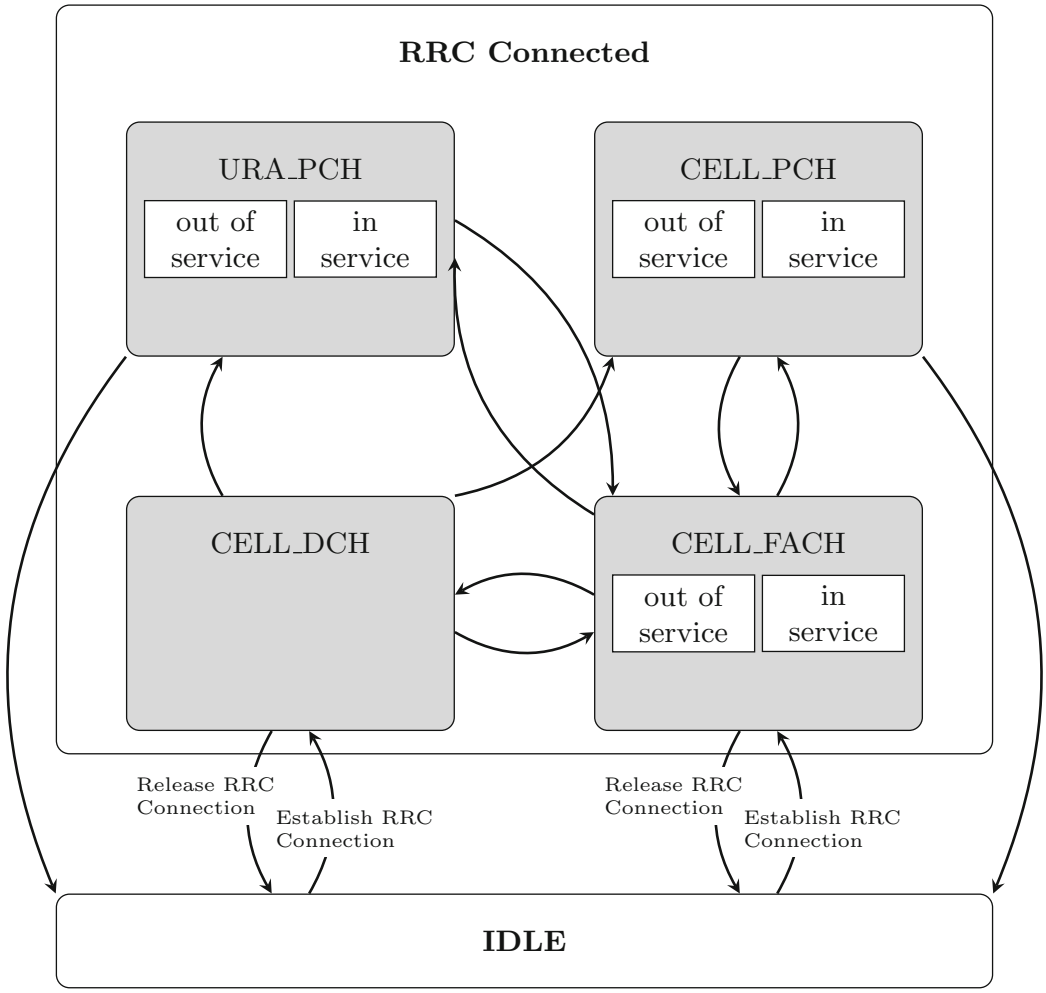


Fig. 6.29 The UMTS RRC states and transitions

PCH. When there is an incoming call or message or the device needs to initiate communication, it transitions to the CELL_FACH state. When the device completes its active communication or when there is a loss of coverage, it releases the RRC connection and transitions to the IDLE state.

- In the URA_PCH state, the mobile device is camping on a UTRAN Registration Area (URA) for extended coverage and power savings. This state is very similar to the CELL_PCH state, except that the UE does not execute cell update after cell reselection. Instead it reads the URA identities from the BCH and only updates its location when the URA changes. This further reduces the power consumption but makes the UE only known on the URA level.

6.4.5.3 RRC Functions

1. Broadcast of System Information
2. Paging
3. Initial Cell Selection and Reselection in Idle Mode
4. Establishment, Maintenance, and Release of RRC Connection
5. Control of RBs, TrCHs, and PhCHs
6. Control of Security Functions
7. Integrity Protection of Signaling
8. UE Measurement Reporting and Control
9. RRC Connection Mobility Functions
10. Support for SRNS Relocation

11. Support for DL Outer Loop Power Control
12. Open-Loop Power Control
13. CBS-Related Functions
14. UE Positioning-Related Functions

6.5 Security

The security approaches of UMTS are based on what was implemented in GSM, including the subscriber authentication, the confidential subscriber identity, the SIM, and the air interface encryption. In addition to the legacy approaches, some new security functions are added and some existing enhanced. Compared to GSM, the encryption algorithm in UMTS is stronger and included also in the Iub interface between Node B and RNC, the application of authentication algorithms is stricter, and the subscriber confidentially is tighter.

6.5.1 UMTS Encryption Algorithm and UMTS Integrity Algorithm

The algorithms introduced by UMTS to protect confidentiality and integrity are f8 and f9, a.k.a. UMTS Encryption Algorithm (UEA) and UMTS Integrity Algorithm (UIA), respectively. They are used to replace the corresponding GSM A5 algorithms. The first specification of f8/f9 is UEA1/UIA1, a.k.a. Kasumi. Later, a second specification UEA2/UIA2, a.k.a. SNOW 3G, was introduced to provide further security enhancements. Note that the UEA can also be configured to UEA0, where no encryption is applied. In contrast, there is no UIA0, i.e., the data integrity protection cannot be disabled.

The f9 algorithm, as shown in Fig. 6.30, derives the message authentication code MAC-I (on the sender side) or XMAC-I (on the receiver side) for signaling messages, based on the following input parameters:

- The 32-bit integrity sequence number COUNT-I, composed of a 4-bit RRC sequence number and a 28-bit RRC hyper frame number.
- The 128-bit integrity key IK, which is established during the authentication and key agreement (AKA) procedure. On the subscriber side, it is stored in the USIM with a copy stored in the ME. On the network side, it is generated in the HLR/AuC, sent to the VLR/SGSN, stored there, and forwarded to the RNC. The IK will expire after a time and shall be regenerated via a new authentication procedure thereafter. It shall be deleted from the memory of ME after power-off or removal of the USIM.
- The 32-bit network-side nonce FRESH, which is dedicated to each user to protect the network against replay of signaling messages by the same user(s).
- The 1-bit direction identifier DIRECTION, which is 0 for UL messages and 1 for DL ones.
- The signaling message itself.

The MAC-I is attached to the sent message and compared on the receiver side with the XMAC-I that is generated from the same message, so as to verify the message integrity.

The f8 ciphering algorithm, as shown in Fig. 6.31, encrypts plaintext by adding a binary keystream bit-per-bit thereto. The plaintext may be recovered by generating the same keystream using the same input parameters and applying a bit-per-bit binary addition with the ciphertext. The generation of keystream is upon the following input parameters:

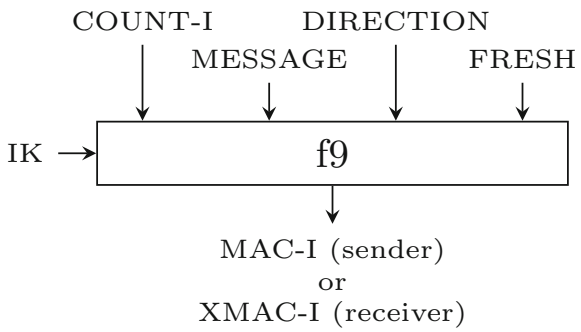


Fig. 6.30 Derivation of MAC-I (or XMAC-I) on a signaling message

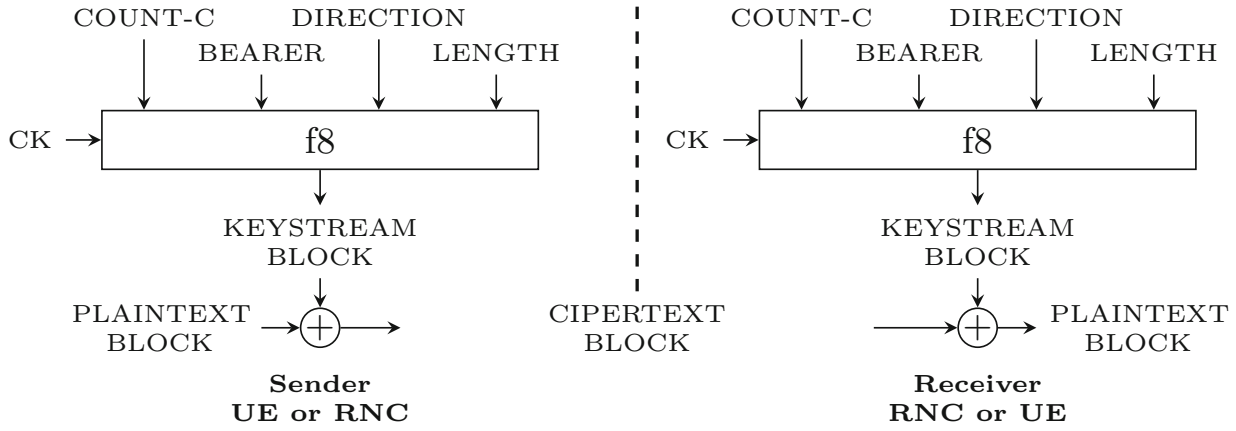


Fig. 6.31 Ciphering of data transmitted over the radio access link

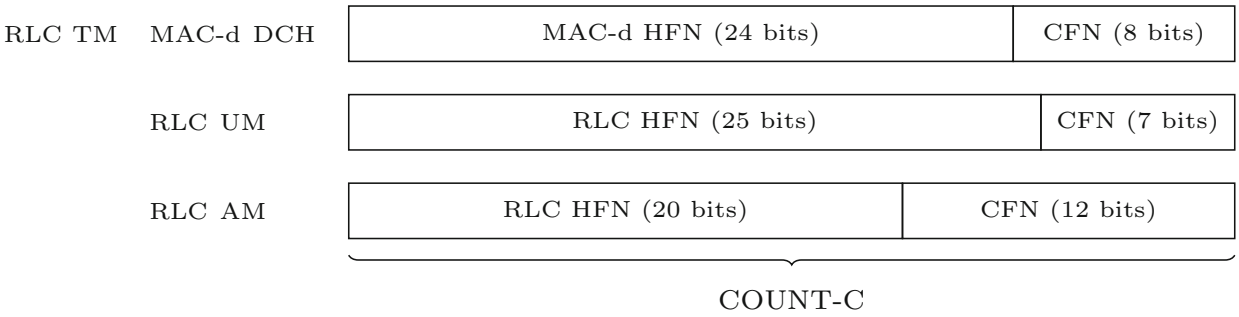


Fig. 6.32 Structure of the ciphering sequence number COUNT-C

- The 32-bit ciphering sequence number COUNT-C, composed of a "short" sequence number and a "long" one, depending on the RLC transmission mode (see Fig. 6.32). There are one COUNT-C value per UL radio bearer and one COUNT-C value per DL radio bearer using RLC AM or RLC UM. For all transparent mode RLC radio bearers of the same CN domain, regardless if UL or DL, the COUNT-C is the same.
- The 128-bit cipher key, which is established during the AKA procedure. There may be one CK for circuit switched (CS) connections and another for PS connections. On the subscriber side, CK is stored in the USIM with a copy stored in the ME. On the network side, it is generated in the HLR/AuC, sent to the VLR/SGSN, stored there, and forwarded to the RNC. The CK will expire after a time and shall be regenerated via a new authentication procedure thereafter. It shall be deleted from the memory of ME after power-off or removal of the USIM.
- The 5-bit radio bearer identifier BEARER, which is to identify different radio bearers associated with the same user and multiplexed on a single PHY layer frame.
- The 1-bit direction identifier DIRECTION like that used in the f9 algorithm.
- The 16-bit length indicator LENGTH that determines the length of the required keystream block.

6.5.2 Mutual Authentication

Though protected by an effective authentication procedure from false subscribers, GSM/GPRS systems have no mechanism to address the threat of false base stations. As a counter measure, the UMTS AKA procedure requires mutual authentication between the subscriber and the network, as briefly illustrated in Fig. 6.33. Upon receipt of an authentication request from the VLR/SGSN, the Home Environment (HE)/AuC sends an array of authentication vectors to the VLR/SGSN, containing random numbers, expected responses, cipher keys, integrity keys, and authentication tokens. Each vector is used for one authentication and key agreement. When initiating authentication, the VLR/SGSN selects the next vector

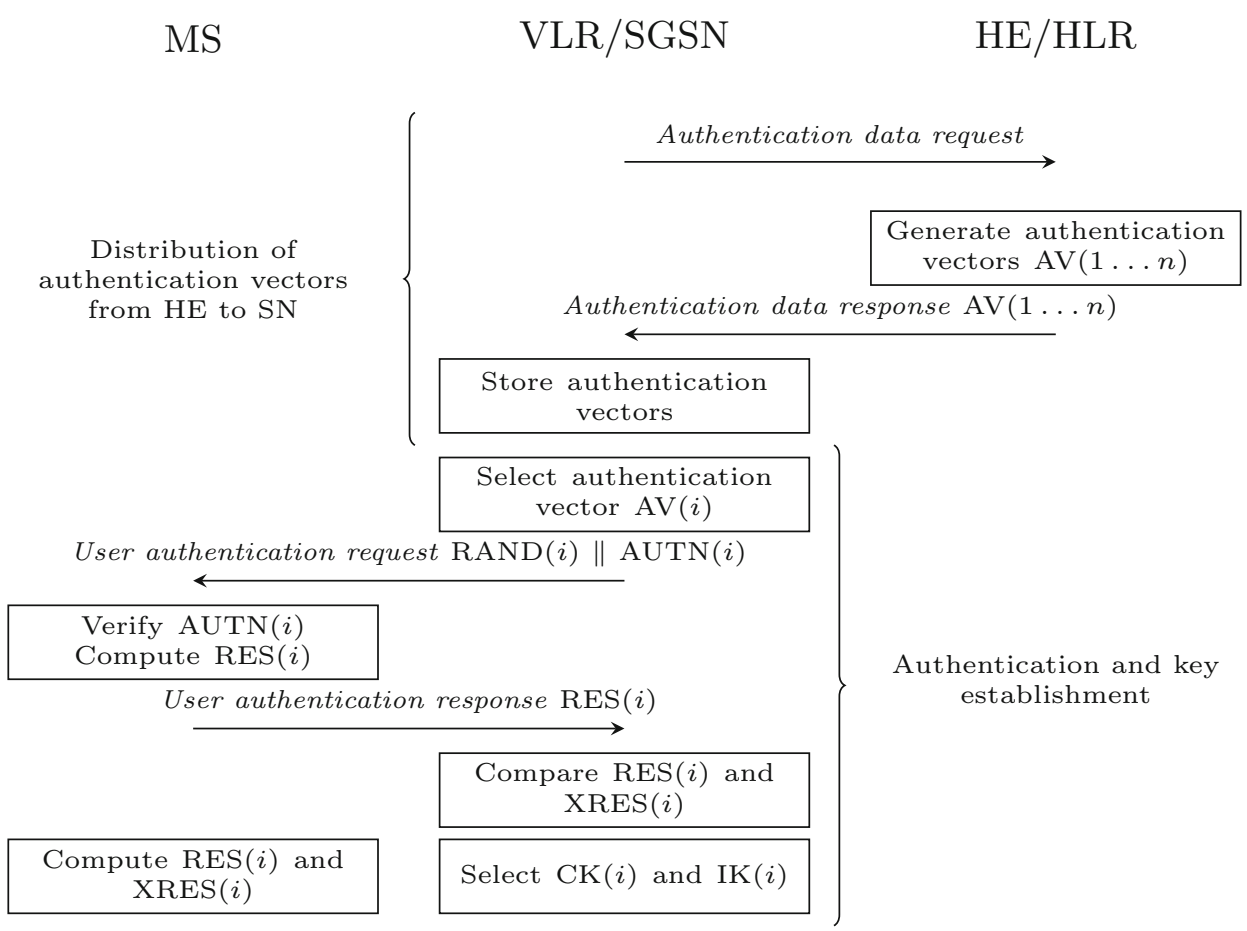


Fig. 6.33 The UMTS AKA procedure

and sends the parameters to the user. The USIM checks the parameters and sends back a response. If the response matches, the authentication and key agreement is successful. The established keys are then transferred for ciphering and integrity functions. In the absence of HE/AuC links, VLR/SGSN can still offer secure service using previously derived keys. Authentication is based on shared integrity keys, ensuring secure connections without the need for reauthentication. Procedures are in place for distributing authentication information, mutual authentication, and key establishment between VLR/SGSN and the mobile station. Additionally, authentication data can be distributed from previously visited to newly visited VLRs, assuming secure links between VLR/Serving GPRS Support Nodes (SGSNs).

6.5.3 User Identity Confidentiality

In UMTS, the user identity confidentiality is safeguarded in three aspects:

1. The IMSI of a user receiving services cannot be intercepted on the radio access link, ensuring its confidentiality.
2. The presence or arrival of a user in a specific area remains undisclosed through eavesdropping on the radio access link, maintaining location confidentiality.
3. By eavesdropping on the radio access link, it is not possible for an unauthorized party to determine if different services are being delivered to the same user, ensuring user untraceability.

6.6 Call Control and Mobility Management

Compared to GSM, UMTS introduces significant advancements in call control and mobility management. These enhancements are designed to address specific challenges and improve the overall user experience.

6.6.1 Handover

The conventional handover procedure in GSM is essentially taken over by UMTS, known as the hard handover. In addition to that, serving the purpose of ensuring seamless connectivity during mobility scenarios, UMTS introduced the so-called *soft handover*. It allows a mobile device to be simultaneously connected to multiple base stations. This feature improves call quality, reduces call drops, and enhances coverage by utilizing signals from multiple sources. It ensures uninterrupted communication even in areas with overlapping coverage. The soft handover algorithm, as illustrated in Fig. 6.34, consists of three measurement-triggered events:

- Radio Link Addition: If $Pilot\_E_c/I_0 > Best\_Pilot\_E_c/I_0 - Reporting\_range + Hysteresis\_event1A$ for a period of ΔT and the active set (the cells that form a soft handover connection) is not full, the measured cell is added to the active set.
- Radio Link Removal: If $Pilot\_E_c/I_0 < Best\_Pilot\_E_c/I_0 - Reporting\_range - Hysteresis\_event1B$ for a period of ΔT, the measured cell is added to the active set.
- Combined Radio Link Addition and Removal: If the active set is full, and $Best\_candidate\_Pilot\_E_c/I_0 > Worst\_Old\_Pilot\_E_c/I_0 + Hysteresis\_event1C$ for a period of ΔT, and the active set (the cells that form a soft handover connection) is not full, the weakest cell in the active set is replaced by the strongest cell in the monitored set.

Note that for the example in Fig. 6.34, the maximum size of the active set is set to 2.

One special form of the soft handover may occur, when a UE can hear the signals from two sectors served by the same Node B. This may occur as a result of the sectors overlapping or more commonly as a result of multipath propagation resulting from reflections from buildings, etc. In this case, multiple radio links from the same Node B can be added to the active set simultaneously, which is known as the *softer handover*.

Designed to be backward compatible and coexisting with legacy systems, UMTS enables handovers between different radio access technologies (RATs), such as GSM and UMTS. This capability facilitates smooth transitions when a mobile device moves between areas covered by different networks. It ensures uninterrupted services and expands the reach of mobile communication.

Furthermore, unlike GSM that operates on a single specific frequency band, UMTS was designed to operate in multiple bands. Therefore, it introduces support to handovers between different carrier frequencies within the UMTS network. This feature is particularly useful in scenarios where different frequencies offer varying coverage or capacity. It enables seamless mobility within the UMTS network by automatically transitioning to the most suitable frequency.

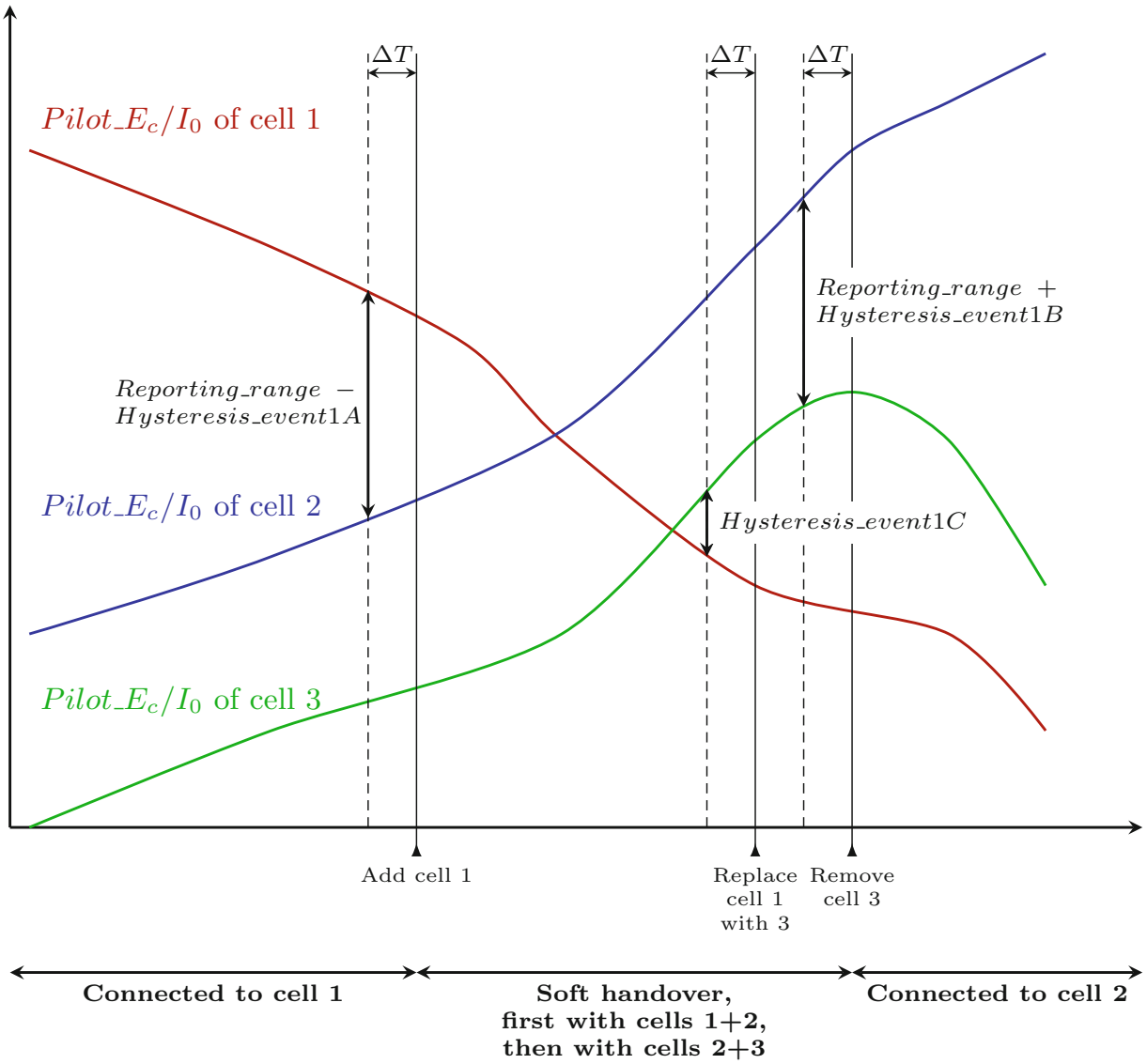
6.6.2 Call Setup and Release Procedures

While GSM relies on the CCCH for call setup and release, UMTS utilizes the DCCH in these procedures. The use of DCCH streamlines the call setup process, resulting in faster call establishment and more efficient allocation of network resources, and therewith improves the user experience.

6.6.3 Supplementary Services

UMTS offers a range of supplementary services to enhance the calling experience and provide users with additional control over their communication. Some of the key services include:

- Call forwarding, which allows users to redirect incoming calls to another number or voicemail. This feature ensures that users do not miss important calls, even when they are unreachable or busy. It provides flexibility in managing incoming calls and ensures effective communication.



$Pilot\_E_c/I_0$	measured and filtered connection quality
$Best\_Pilot\_E_c/I_0$	strongest $Pilot\_E_c/I_0$ of all cells in the active set
$Best\_candidate\_Pilot\_E_c/I_0$	strongest $Pilot\_E_c/I_0$ of all cells in the monitored set
$Worst\_Old\_Pilot\_E_c/I_0$	weakest $Pilot\_E_c/I_0$ of all cells in the active set
$Reporting\_range$	threshold for soft handover
$Hysteresis\_event1A$	addition hysteresis
$Hysteresis\_event1B$	removal hysteresis
$Hysteresis\_event1C$	replacement hysteresis
ΔT	Time to trigger

Fig. 6.34 The soft handover procedure

- Call waiting, which allows users to receive notifications and switch between incoming calls while they are already on a call. This feature enables efficient call management and prevents missed calls, allowing users to handle multiple calls seamlessly.
- Call barring, which allows users to restrict certain types of outgoing or incoming calls based on specific criteria. This feature offers control over call usage, helps manage costs, and provides privacy by blocking unwanted calls.

- Multi-numbering, which facilitates the use of multiple numbers on a single SIM card, allowing users to have different phone numbers for personal and business use. This feature simplifies call management and provides flexibility in communication.

By incorporating these features and enhancements, UMTS significantly improves call control, mobility management, and the overall user experience, offering users greater flexibility, seamless connectivity, and efficient call management.

6.7 Location Service

Location services in UMTS provide the capability to determine the geographical location of a mobile device within the network. This information can be used for various purposes, such as emergency services, location-based services, and network optimization. In UMTS, location services are enhanced compared to GSM, offering more precise and accurate location determination.

6.7.1 Location Services Categories

UMTS introduces four categories of location services (LCS):

- **Commercial LCS:** The Commercial LCS, also known as Value-Added Services, is associated with applications that provide value-added services to subscribers based on their location information. These services utilize the knowledge of the mobile device's location (and optionally velocity) and, if available and authorized by the operator, the positioning method used to obtain the location estimate. Examples of Commercial LCS include restaurant directories with directions based on the user's current location.
- **Internal LCS:** The Internal LCSs is designed to leverage the location information of mobile devices for Access Network internal operations. It includes functionalities such as location-assisted handover, traffic and coverage measurement, support for operation and management tasks, supplementary services, and integration with GSM bearer services and teleservices. The Internal LCS enhances network performance and optimization by utilizing accurate location information.
- **Emergent LCS:** The Emergency LCS is a critical service provided to assist subscribers who make emergency calls. When emergency calls are placed, the location of the mobile device caller and, if available and authorized, the positioning method used to obtain the location estimate are provided to the emergency service provider. This information aids emergency responders in their response efforts. In some jurisdictions, such as the United States, the Emergency LCS is mandated for all mobile voice subscribers.
- **Lawful Intercept LCS:** The Lawful Intercept LCS utilizes location information to support legally required or sanctioned services. It assists law enforcement agencies or authorized entities in accessing and monitoring the location of specific mobile devices for lawful interception purposes. This service ensures compliance with legal and regulatory requirements related to interception and surveillance activities.

In general, UMTS expands on the location services categories introduced in GSM, enabling more sophisticated and specialized applications based on accurate and reliable location information. The different categories of location services cater to various user needs, ranging from value-added services to emergency response support and legal interception requirements. These enhancements in UMTS enable the network to deliver a wide range of location-based services and facilitate efficient operations within the RAN.

6.7.2 Positioning Methods

UTRA supports various LCS methods, including (i) cell coverage-based positioning method, (ii) Observed Time Difference of Arrival (OTDOA) positioning method, (iii) Assisted Global Navigation Satellite System (A-GNSS)-based positioning methods, and (iv) Uplink Time Difference of Arrival (U-TDOA) positioning method.

Table 6.8 Summary of functional groups and functional blocks for UMTS LCS

Functional group	Functional component	Functional block	Abbreviation
Location Client	Location client component	(External) Location Client Function	LCF
		Location Client Function	LCF-internal
LCS Server in PLMN	Client handling Component	Location Client Control Function	LCCF
		Location Client Authorization Function	LCAF
		Location Client Co-ordinate Transformation Function	LCCTF
		Location Client Zone Transformation Function	LCZTF
	System handling component	Location System Control Function	LSCF
		Location System Billing Function	LSBF
		Location System Operations Function	LSOF
		Location System Broadcast Function	LSBcF
		Location System Co-ordinate Transformation Function	LSCTF
		Location IMS—Interworking Function	LIMS-IWF
	Subscriber handling component	Location Subscriber Authorization Function	LSAF
		Location Subscriber Translation Function	LSTF
		Location Subscriber Privacy Function	LSPF
	Positioning component	Positioning Radio Control Function	PRCF
		Positioning Calculation Function	PCF
		Positioning Signal Measurement Function	PSMF
		Positioning Radio Resource Management	PRRM

6.7.3 LCS Architecture

The UMTS LCS is implemented with a set of different functional blocks, which can be categorized into several functional components and further into two functional groups, as summarized in Table 6.8. The allocation of these LCS functional entities to network elements is shown in Fig. 6.35.

6.7.4 Signaling and Interfaces

In UMTS, the LCS signaling between RAN and CN is accomplished through the Iu interface—mainly involved procedures are the Location Request by CN and Location Report by UE. The newly introduced Gateway Mobile Location Centre (GMLC) is connected to the HSS, the MSC/SGSN, and the GSM's Customized Applications for Mobile networks Enhanced Logic (CAMEL) over the Lh, Lg, and Lc interfaces, respectively, which are cumulatively known as the Mobile Application Part (MAP) interfaces. The MAP services include (i) sending routing information, (ii) providing subscriber location, and (iii) reporting subscriber location. The Lpp interface between the GMLC and the Privacy Profile Register (PPR) is responsible for the request and response of LCS authorization, as well as the LCS privacy profile update notification and the acknowledge thereto. Last but not least, the Lid interface between the GMLC and the Pseudonym Mediation Device (PMD) is used to exchange LCS identity request and LCS identity response.

6.7.5 LCS States

The LCS states are defined in the GMLC differently from that in the MSC server and SGSN, as shown in Fig. 6.36.

6.8 IP Multimedia Subsystem/IMS

To enable the delivery of multimedia services over IP networks, UMTS introduces a new fundamental architectural framework, known as the IMS. In UMTS networks, IMS plays a crucial role in supporting advanced multimedia

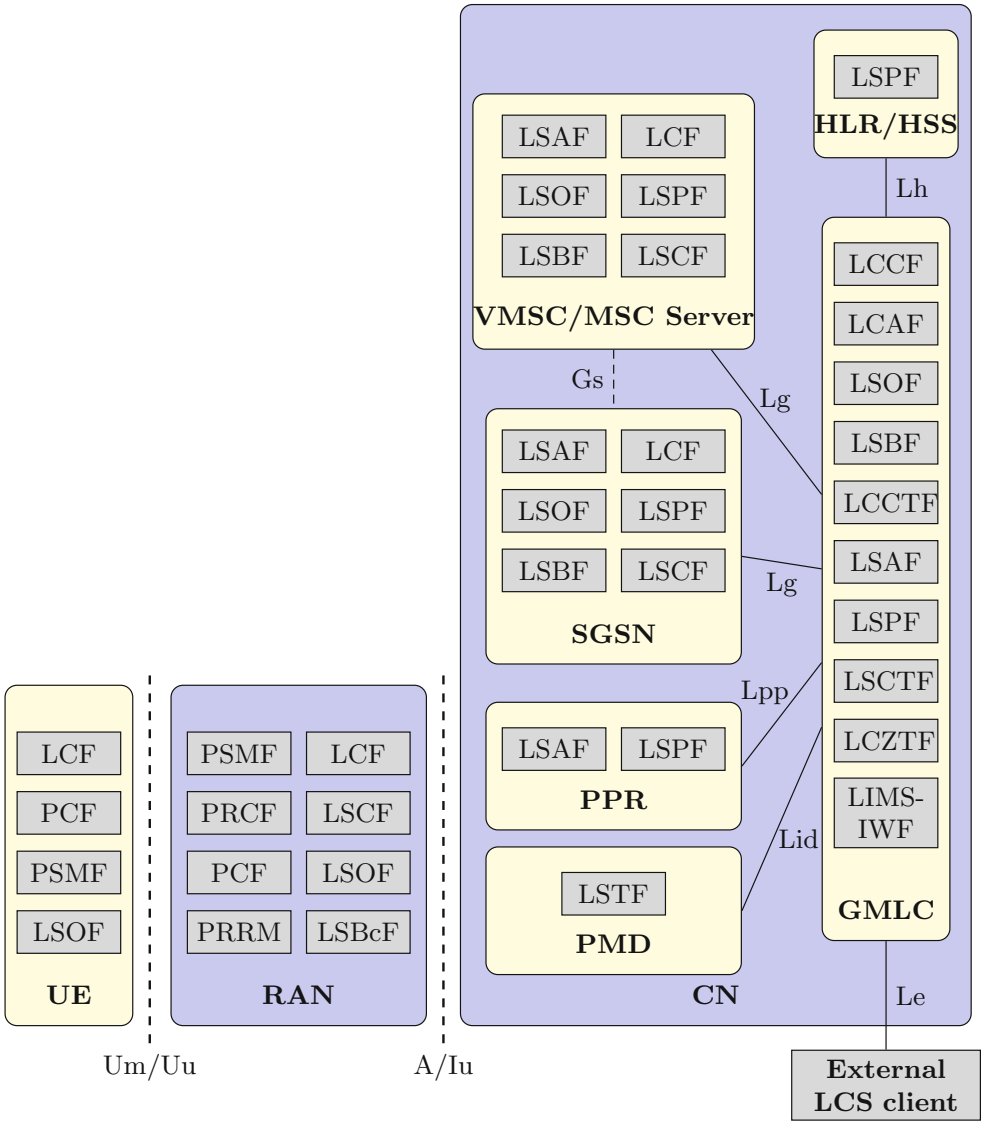


Fig. 6.35 The generic LCS logical architecture and interfaces

communication and services. It provides a flexible and scalable platform for delivering services such as VoIP, video calling, instant messaging, presence, and multimedia conferencing. IMS represents a shift toward IP-based networks and facilitates the convergence of communication services.

The architecture of IMS, as illustrated in Fig. 6.37, consists of several key functional entities that work together to enable multimedia services. At the core of IMS is the Call Session Control Function (CSCF), which serves as the main signaling and control element. It includes three subcomponents: Proxy-CSCF (P-CSCF), Serving-CSCF (S-CSCF), and Interrogating-CSCF (I-CSCF). The Media Gateway Control Function (MGCF) handles the interconnection between IMS and legacy networks. The Media Gateway (MGW) enables the conversion of media streams between different network types. Application Serverss (ASs) provide specific services and applications to users, while the HSS stores user-related information.

IMS uses the Session Initiation Protocol (SIP) as the primary signaling protocol for session establishment, modification, and termination. SIP enables the control and management of multimedia sessions in IMS. The interfaces between different IMS components are essential for the exchange of signaling and media information. These interfaces include the interfaces between CSCFs, MGCFs, and MGWs. The IMS control plane handles the signaling and control functions, while the user plane carries the actual media streams.

IMS enables a wide range of multimedia services and applications. These include voice and video communication, instant messaging, presence information, and multimedia conferencing. IMS facilitates the integration of various communication

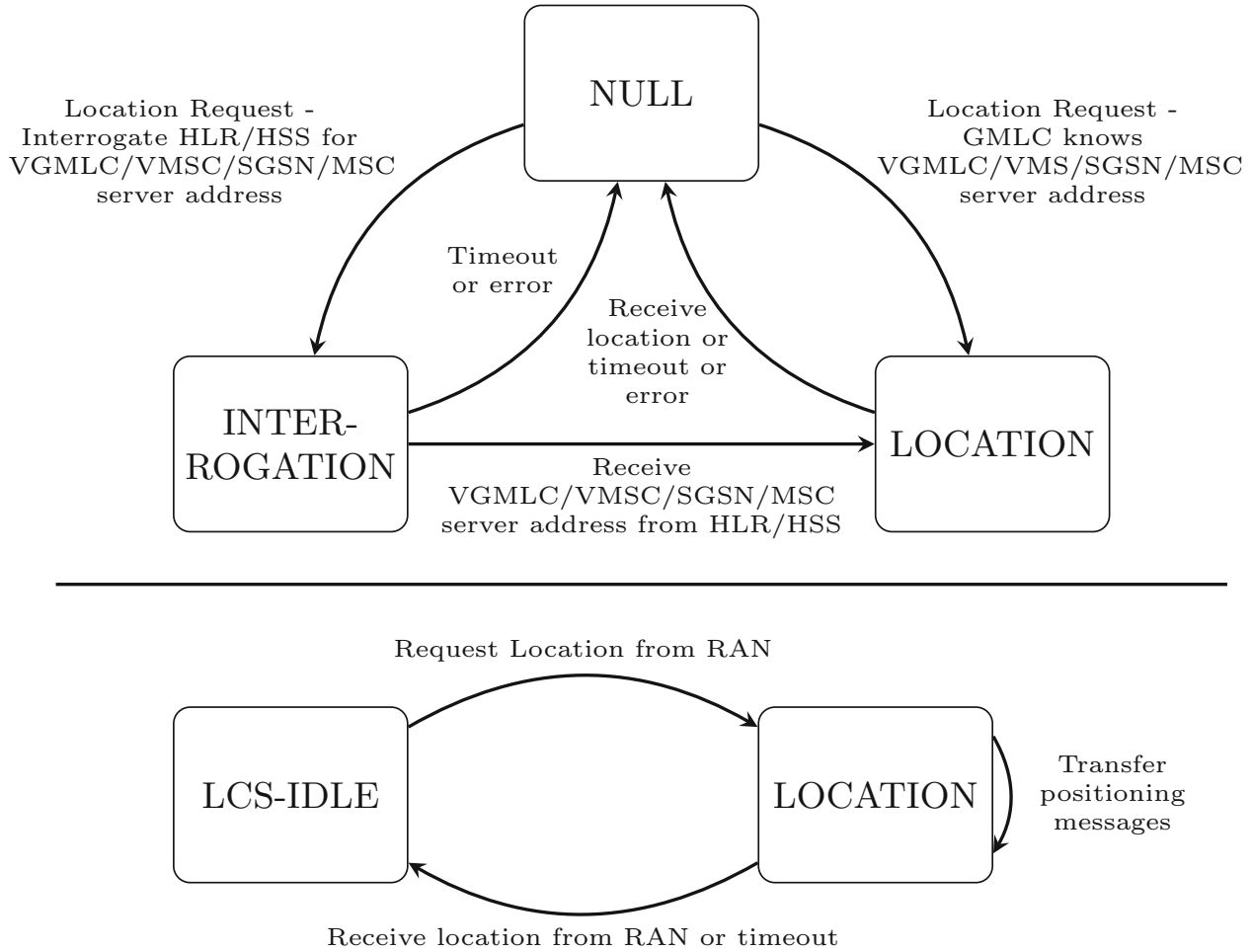


Fig. 6.36 The LCS state transitions in the GMLC (above), as well as in the MSC server and SGSN (below), respectively

services, allowing users to seamlessly access multimedia services across different networks and devices. The flexibility and scalability of IMS enable the rapid deployment and evolution of new multimedia services.

In UMTS networks, IMS is integrated into the network architecture to provide multimedia services and interacts with other UMTS element. The integration of IMS brings several benefits, including service flexibility, scalability, and interoperability. It allows UMTS operators to offer advanced multimedia services and leverage the advantages of IP-based networks.

6.9 Summary

In this chapter, we focused on the UMTS, the most successful 3G standard that played a critical role in the advancement of mobile communications. To offer an insightful view of the operation of a 3G system, this chapter provides an in-depth analysis of UMTS's system architecture, its interfaces, and the technologies that underpin it. We delved into the design considerations and the unique features that set UMTS apart, such as advanced roaming capabilities and support for location-based services. The chapter also examines the challenges and limitations of UMTS, offering insights into how it fits into the broader context of mobile communications evolution. Readers will come away with a thorough understanding of UMTS, its significance in the 3G era, and its contributions to the ongoing development of mobile technologies.

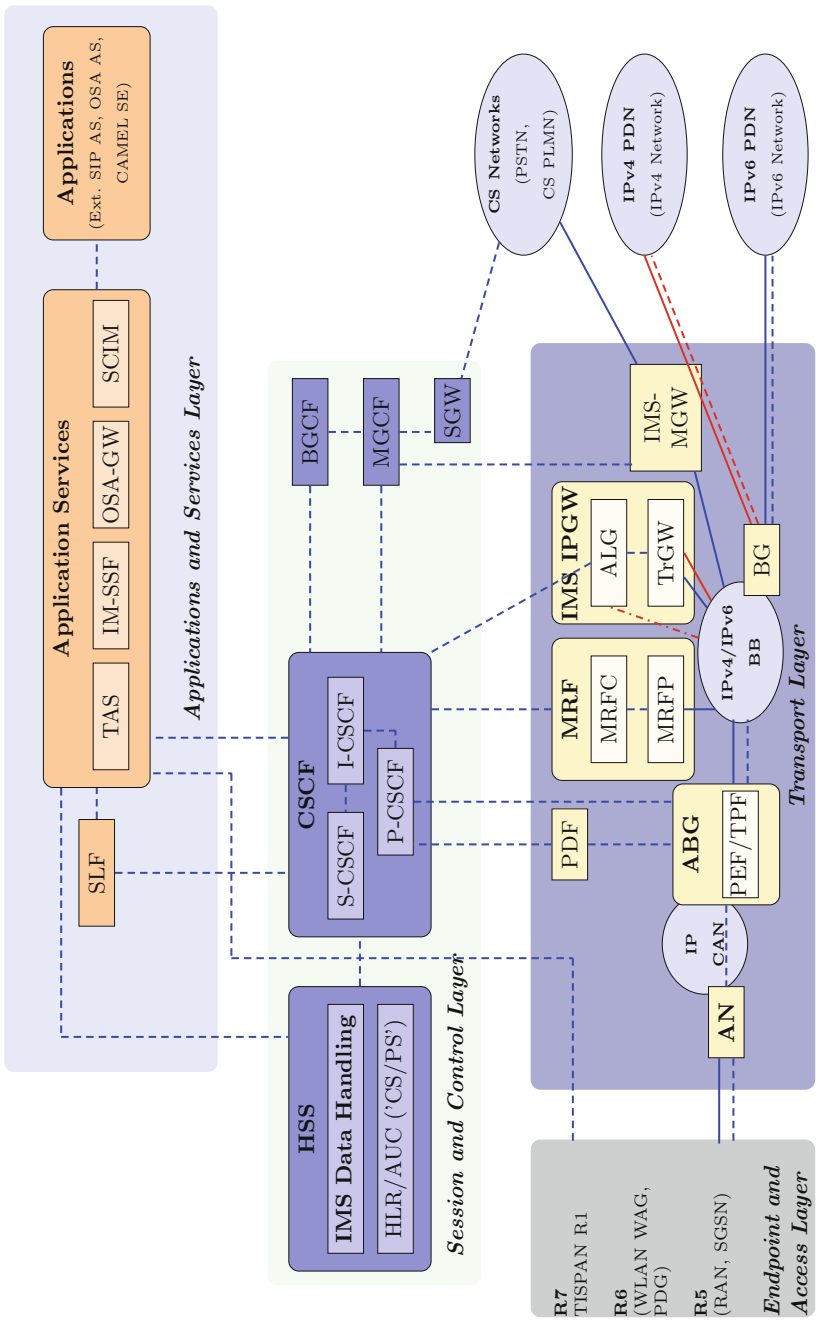


Fig. 6.37 The IMS architecture

6.10 Exercises

1. List the main components of the UMTS core network, and describe their functions.
2. How does the architecture of the UMTS core network differ from that of GSM?
3. Describe the role of Node B and RNC in the UMTS RAN.
4. Discuss the advantages of having a separate control and user plane in UMTS RAN.
5. Explain the terms “hard handover” and “soft handover” in the context of UMTS.
6. Given a mobility scenario, outline the steps involved in a soft handover process in UMTS.
7. List and describe the four categories of LCS in UMTS.
8. Discuss the importance of Emergency LCS and how it could be critical in real-world applications.
9. What are the entities CSCF, MGCF, and MGW, respectively? What are their functionalities?

Evolution to Fourth-Generation (4G) Mobile Cellular Communications

7

7.1 4G: All-IP Mobile Internet

During the first decade of the twenty-first century, there was a significant increase in the number of mobile subscribers. The milestone of reaching one billion subscribers was achieved in 2002, but the number grew rapidly, surpassing five billion in 2010. This exponential growth was fueled by the surge in the volume of mobile broadband traffic, which exceeded that of cellular voice traffic for the first time. However, it imposed a challenge for cellular networks that were primarily optimized for voice communications and had to operate both a packet-switched network for data delivery and a circuit-switched network for voice calls since the data services were introduced in the 2.5G systems. As a result, the 4G cellular system was developed with an end-to-end all-IP architecture where only a packet-switched network was provided to offer Internet-based services more effectively. This design completely abandoned the circuit-switched network, for the first time in the history of cellular systems, and voice service was transferred to a kind of data service known as VoIP.

7.2 IMT-Advanced Cellular Standards

In parallel with the roll-out of commercial 3G systems in the early 2000s, the standardization bodies including 3GPP, 3GPP2, the WiMAX Forum, and IEEE, embarked on continuous efforts to enhance the existing 3G standards and pave the way for 4G technology. To improve the performance and capacity, advanced air interface technologies adapted to the design of all-IP network infrastructure were utilized. Recognizing the need to maintain the competitiveness of UMTS, 3GPP initiated in 2004 the study item of the *Long-Term Evolution* air interface, also referred to as Evolved Universal Terrestrial Radio Access (E-UTRA). This marked the beginning of a significant endeavor toward 4G. In December 2008, the initial version of the LTE air interface and its associated core-network technology called the Evolved Packet Core (EPC) were completely specified in 3GPP Release 8. Following this milestone, the enhanced version of LTE and EPC was specified as the freeze of 3GPP Release 9 in December 2009.

Like the development process of IMT-2000, the ITU-R Working Party 5D (WP5D) released the essential technical requirements for the 4G system in 2008, known as the International Mobile Telecommunications-Advanced (IMT-Advanced), in its report (ITU-R M.2134, 2008). However, it became apparent that LTE did not entirely fulfill the requirements specified by IMT-Advanced. For instance, the expected peak data rates of 1 Gbps for low mobility and 100 Mbps for high mobility were not fully met. To close these gaps, 3GPP continued its efforts and focused on an enhanced version in Release 10 known as Long-Term Evolution Advanced (LTE-Advanced) or LTE-A starting in 2009 (Parikh & Basu, 2011). LTE-Advanced aimed to further augment the capabilities of LTE, bridging the remaining disparities between LTE and the rigorous demands set by IMT-Advanced.

During this period, the WiMAX standard also underwent continuous evolution under the joint efforts of IEEE and the WiMAX Forum. In 2011, the specifications for WirelessMAN-Advanced, also referred to as Mobile WiMAX Release 2.0, were finalized in order to fully conform with the IMT-Advanced requirements. It was a big leap as notable enhancements were introduced to the legacy IEEE802.16e-2005, giving rise to a new standard named IEEE802.16m-2011. IEEE and the WiMAX Forum jointly submitted IEEE802.16m-2011 to the ITU-R as one of the proposals for IMT-Advanced, thereby directly competing with LTE-Advanced as the technological solutions for IMT-Advanced. In parallel, 3GPP2 was actively

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5_7

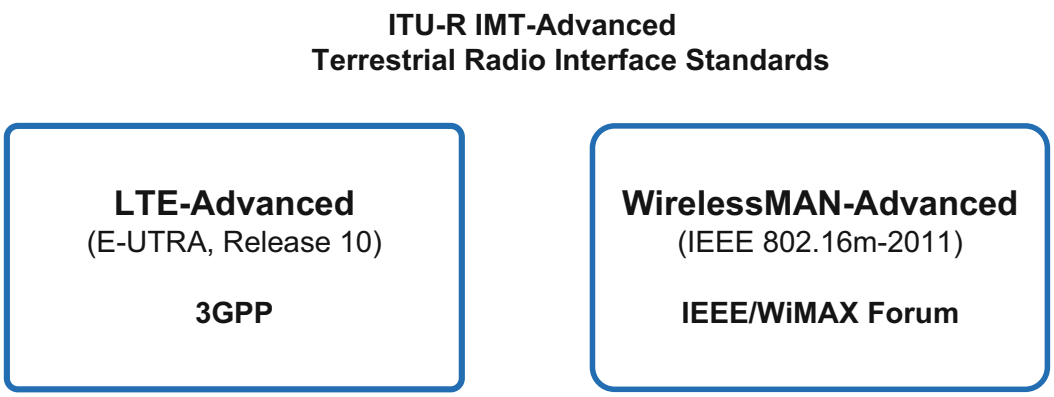


Fig. 7.1 Two IMT-Advanced standards recommended by ITU-R M.2012

engaged in developing UMB as a successor to CDMA2000, positioning it as a next-generation technology. However, Qualcomm, a leading sponsor of UMB, made a significant announcement in November 2008, declaring the cessation of UMB's development in favor of LTE. This decision had far-reaching consequences for 3GPP2, as it marked the end of their activities in 2013. Consequently, the standardization body has remained inactive since then, with no further developments or contributions.

In October 2010, the ITU-R completed the evaluation of six candidate submissions and approved two industry-developed technologies as the global 4G standards, as shown in Fig. 7.1. Recommendation (ITU-R M.2012, 2012) identifies the terrestrial radio interface technologies of IMT-Advanced and provides detailed radio interface specifications.

7.2.1 LTE-Advanced

Along with the smooth evolution of UTRA, which was upgraded from WCDMA, HSDPA, HSUPA, to HSPA+, the 3GPP launched a study item in late 2004 to explore a disruptive radio access technology designed specifically for all-IP data transmission. This study item focused on defining the technical requirements for LTE. In June 2005, several significant features were approved, including low latency, high data rates at the cell edge, and flexibility in spectrum allocation. During the 3GPP RAN plenary meeting in December 2005, a decision was made to utilize orthogonal frequency-division multiple access (OFDMA) for the downlink transmission of LTE and single-carrier frequency-division multiple access (SC-FDMA) for the uplink transmission. The initial version of the LTE air interface (Release 8) and the accompanying EPC core network were finalized in December 2008, followed by an enhanced version (Release 9) that was frozen in December 2009.

LTE is a groundbreaking standard that was intentionally designed without the constraint of backward compatibility, allowing for greater flexibility in adopting advanced technical features. It can be operated in two duplexing modes: LTE FDD and LTE TDD. The latter is also known as Time Division-Long Term Evolution (TD-LTE), which is an evolutionary path from the TD-SCDMA at the initial stage. In terms of data rates, LTE achieves peak speeds of up to 300 Mbps in the downlink and 75 Mbps in the uplink, utilizing a signal bandwidth of 20 MHz. In addition, it offers a flat system architecture to simplify the system design and lower end-to-end latency. In December 2009, TeliaSonera introduced the world's first commercial LTE mobile services in the capitals of Scandinavia, i.e., Stockholm and Oslo. Ericsson and Huawei supplied the network equipment for this milestone launch (Astely et al., 2009). Notably, there were no commercially available LTE-compliant mobile phones at that time, so the early subscribers had to utilize their computers with a USB wireless network adapter to access LTE services. It wasn't until September 2010 that the Samsung SCH-r900, the world's first LTE-compliant mobile phone, was finally released.

Due to LTE's inability to fully meet the requirements of IMT-Advanced, such as the desired peak data rates of 1 Gbps for low-mobility environments and 100 Mbps for high mobility scenarios, 3GPP decided to develop an improved version called LTE-Advanced or LTE-A starting in 2009. The initial proposal for LTE-Advanced was submitted to the ITU-R in October 2009, and more detailed specifications were later finalized to create the first version of LTE-Advanced in Release 10. LTE-Advanced introduced significant technical advancements like enhanced MIMO and wider bandwidth of up to 100 MHz via the use of carrier aggregation (CA), achieving high speed transmission of 1 Gbps in the downlink and 500 Mbps in the

uplink. In 2012, YOTA Networks, a Russian operator, announced the launch of the world's first LTE-Advanced network in Moscow using equipment from Huawei. In early 2016, the completion of 3GPP Release 13 specifications marked the introduction of *LTE-Advanced Pro*. It is an intermediate step between 4G and 5G, designed to provide higher data rates, improved network capacity, lower latency, and better overall user experience. While it does not meet the requirements and capabilities of 5G technology, it offers significant improvements over IMT-Advanced, expanding LTE to a wide range of new industries and enabling new use cases beyond smartphones, such as automotive and IoT. It is noted that both LTE-Advanced and LTE-Advanced Pro maintain backward compatibility with the purpose of smooth upgrading from LTE (Dahlman et al., 2021).

To provide readers with a comprehensive view, we briefly summarize the evolution of the 3GPP LTE releases below, as also illustrated in Fig. 7.2.

- *Release 8* was the first definition of the LTE radio access technology and the all-IP EPC network, forming the foundation for the following evolution. Spectrum flexibility was emphasized by supporting paired and unpaired spectrum using FDD and TDD, respectively. It also supports flexible bandwidths (i.e., 1.4 MHz, 3 MHz, 5 MHz, 10 MHz, 15 MHz, and 20 MHz) by scaling the number of OFDM subcarriers. Over a bandwidth of 20 MHz, the peak data rate can reach 150 Mbps in the uplink with 2×2 MIMO and 300 Mbps in the downlink with 4×4 MIMO.

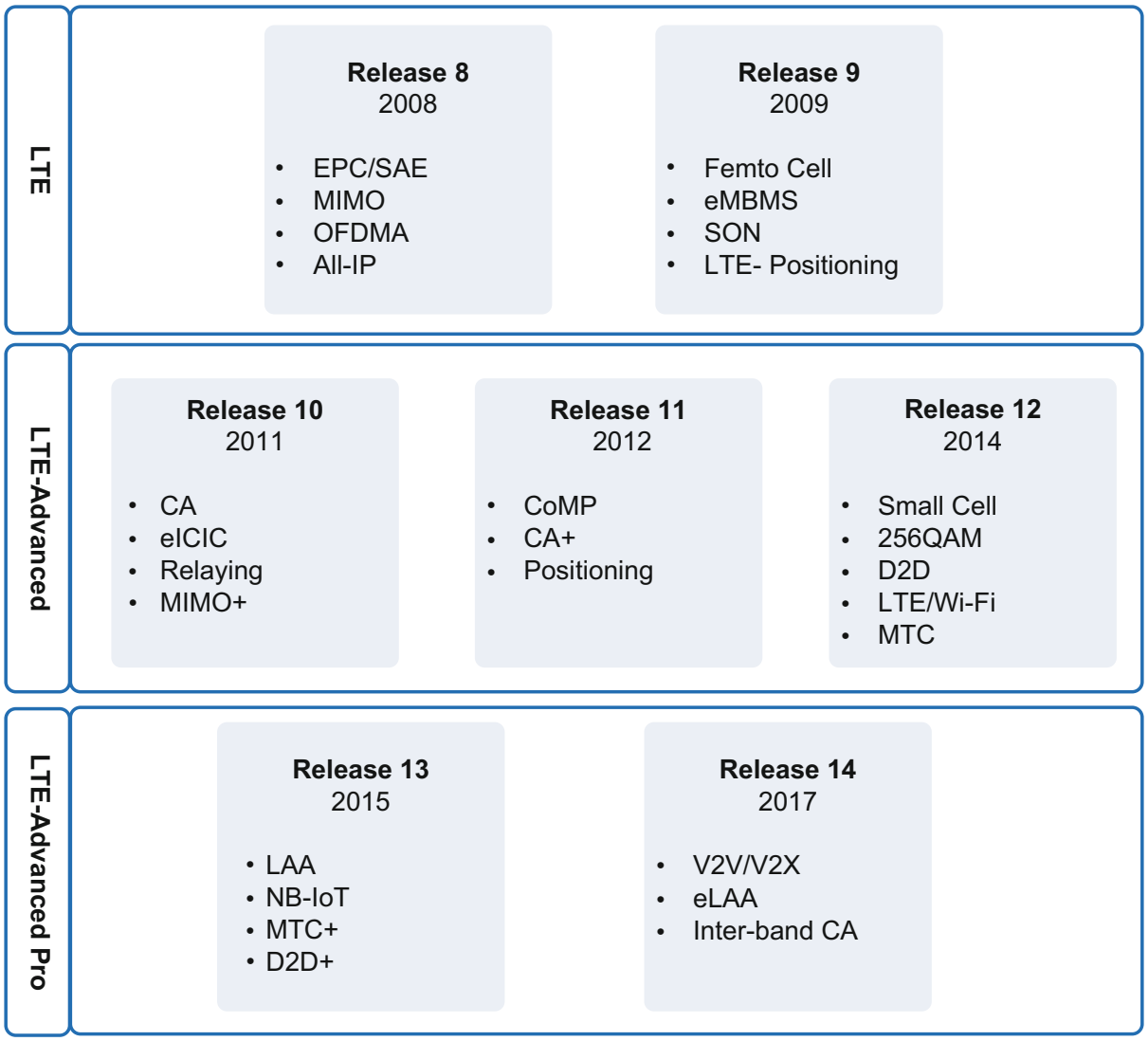


Fig. 7.2 The evolution of 3GPP LTE releases

- *Release 9* was the first step of LTE evolution. It provided some improvements left behind from Release 8 and brought several advancements to LTE, including femtocell, MIMO beamforming, self-organizing networks (SON), enhanced Multimedia Broadcast Multicast Services (eMBMS), LTE positioning, and a public warning system. These features improved indoor connectivity, increased network capacity, automated network management, facilitated efficient multimedia broadcasting, empowered location-based services, and enabled critical emergency notifications.
- *Release 10* focused on the specification of LTE-Advanced, aiming to ensure full compliance with the IMT-Advanced requirements for LTE radio access technology. Frozen in March 2011, this release introduced many technical features such as carrier aggregation, enhanced uplink multiple access, MIMO enhancements, relaying, enhanced ICIC (eICIC), heterogeneous network deployment, and SON enhancements.
- *Release 11* was finalized in September 2012, which brought improvements to the performance and capabilities of LTE-Advanced. One standout feature of Release 11 was the introduction of coordinated multi-point transmission and reception (CoMP), which revolutionized communication by enabling cooperation between multiple base stations. Alongside CoMP, Release 11 also featured enhancements in carrier aggregation, a new control-channel structure, network-based positioning, RAN overload control for machine-type communication, and a smartphone battery-saving technique. These advancements propelled LTE-Advanced to new heights, delivering better network performance, increased efficiency, enhanced positioning capabilities, optimized resource allocation, and improved battery life for smartphones.
- *Release 12* was completed in June 2014 and focused on the optimization and enhancements for small cells, including dual connectivity, dense small-cell deployment, small-cell on/off, and semi-dynamic TDD. Higher-order modulation (i.e., 256QAM) was introduced to make use of high signal strength in a small-cell environment. Another priority of this release was applying LTE technology for emergency events and public safety, with technical specifications for mission-critical application layer functional elements. Other features included device-to-device (D2D) communications, LTE TDD-FDD joint operation including carrier aggregation, security assurance methodology, and LTE-Wi-Fi integration.
- *Release 13* marked the start of LTE-Advanced Pro, which was sometimes in marketing dubbed 4.5G and seen as an intermediate step between the first release of LTE and the advent of 5G. Release 13 was a significant leap, with many interesting features, such as license-assisted access (LAA) to support unlicensed spectra, improved support for machine-type communication (MTC), and further enhancements in MIMO, D2D communications, and carrier aggregation. Other efforts for expanding it to a set of new services and new verticals included the introduction of NB-IoT and the initial studies on vehicle-to-vehicle (V2V) communications.
- *Release 14* was frozen in 2017. In addition to improving previously introduced features like enhanced license-assisted access (eLAA) and inter-band carrier aggregation, Release 14 introduced support for V2V and vehicle-to-everything (V2V) communications. It also brought wide-area broadcast capabilities with reduced subcarrier spacing. These advancements expanded the capabilities of LTE, enabling seamless communication between vehicles and various connected devices, while also enhancing the efficiency of wide-area broadcasts with improved subcarrier spacing.

7.2.2 WirelessMAN-Advanced

In 2007, the ITU-R granted approval to WiMAX Release 1.0, formally known as IMT-2000 OFDMA TDD WMAN, as one of the global 3G standards. As previously mentioned, WiMAX Release 1.0 was based on the transmission algorithms and access protocols specified in IEEE 802.16e-2005. After the approval of WiMAX Release 1.0, the IEEE WMAN community recognized the need to further evolve the WiMAX technology and therefore developed an advanced version referred to as IEEE 802.16m, aiming at boosting the performance and capabilities of WiMAX technologies to ensure full compliance with the requirements of IMT-Advanced defined by ITU-R M.2134 (2008). Evaluation results indicated that LTE-Advanced and IEEE 802.16m exhibit comparable performance, which was not surprising considering their adoption of similar cutting-edge technologies during that time. As a result, IEEE 802.16m, similar to LTE-Advanced, satisfied all the IMT-Advanced requirements. In October 2010, the ITU-R released its recommendation (ITU-R M.2012, 2012), where WiMAX Release 2.0 (built on IEEE 802.16m) under the name of WirelessMAN-Advanced, along with 3GPP LTE-Advanced, were identified as two IMT-Advanced standards.

Unlike LTE-Advanced, which is a backward compatible evolution of LTE, IEEE 802.16m took a different approach and introduced some disruptive features (Dahlman et al., 2011). It was recognized as a new standard rather than a smooth evolution of IEEE 802.16e, although it retained certain essential characteristics from IEEE 802.16e, such as the basic OFDM numerology. Both IEEE 802.16e and IEEE 802.16m can coexist on the same carrier within the 5 ms frame structure through time multiplexing. IEEE 802.16m incorporated several features similar to those of LTE-Advanced, including the adoption of

carrier aggregation for wide bandwidths exceeding 20 MHz and support for relaying functionality. It also introduced shorter subframes, approximately 0.6 ms in length, to reduce the round-trip time of hybrid ARQ and overall latency in the radio interface. Instead of inheriting the resource-mapping schemes defined in its predecessor, IEEE 802.16m introduced physical resource units consisting of a set of frequency-contiguous subcarriers within one subframe. This approach is similar to the design of resource blocks in LTE. Each resource unit in IEEE 802.16m consists of 18 subcarriers with an inter-subcarrier spacing of 10.94 kHz, resulting in a bandwidth that is close to the LTE resource-block bandwidth of 180 kHz.

WiMAX was commercialized earlier than LTE and enjoyed a period of superiority in terms of data throughput from 2005 to 2009 (Etemad, 2008). It was at the forefront of adopting advanced wireless technologies like MIMO and OFDM. Additionally, WiMAX introduced innovative features such as variable transmission bandwidths. In 2006, two South Korean telecom operators launched the world's first mobile WiMAX service based on the IEEE 802.16e standard, branded as WiBro (Wireless Broadband). The WiMAX Forum reported the deployment of around 600 WiMAX networks (offering either fixed access or mobile broadband) in more than 148 countries by the end of 2010, covering a population of over 621 million people. However, LTE had the unique advantage of evolving from dominant standards like GSM and WCDMA, while WiMAX was considered a more disruptive technology with a smaller user base. Consequently, major mobile operators such as Verizon, Vodafone, China Mobile, NTT, and Deutsche Telekom opted to smoothly upgrade their existing 3G infrastructure to LTE rather than undertake the risk and high cost of using a disruptive standard.

In the end, LTE/LTE-Advanced won the competition to become the dominant 4G standard. With LTE/LTE-Advanced, the world has thus converged into a universal global standard for mobile communications, deployed by most of mobile network operators worldwide and applicable to both paired and unpaired spectra.

7.3 Key Technologies for 4G

The development of 4G has two key requirements: high spectral efficiency and spectrum flexibility. To meet these requirements, both LTE and WiMAX adopted MIMO and OFDM, also jointly referred to as MIMO-OFDM (Bolcskei, 2006), as the fundamental transmission scheme. In order to address emerging needs while maintaining backward compatibility, 3GPP and IEEE continuously integrated new technical features into the initial standards. These enhancements aimed to ensure full compliance with the IMT-Advanced requirements. Using LTE-Advanced as the example, its Release 10 improved spectrum flexibility through CA, enhanced multi-antenna transmission, and introduced support for relaying and heterogeneous network deployment. Release 11, which was finalized in late 2012, further enhanced the capabilities of LTE. The most notable feature of this release was the enhancements of radio interface functionalities for CoMP transmission and reception. Release 12, which was completed in 2014, introduced new scenarios by enabling D2D communications and supporting low-complexity devices for machine-type communications. With the advent of Release 13, a significant evolution called LTE-Advanced Pro emerged. This release incorporated LAA to support unlicensed spectra alongside licensed spectra, thereby expanding the available resources (Astely et al., 2009).

In the rest of this chapter, we will introduce the fundamentals of key technological enablers that have played a vital role in the development and evolution of 4G.

7.3.1 Multi-Input Multi-Output/MIMO

The impressive potential demonstrated by the vertical Bell Laboratories layered space-time architecture (V-BLAST) multi-antenna wireless system in the late 1990s attracted significant interest in MIMO technologies (Foschini, 1996). By utilizing multiple antennas, MIMO allows the exploitation of spatial dimensions as an additional degree of freedom for wireless communications, achieving high spectral efficiency. In environments with rich scattering, the theoretical spectral efficiency scales linearly with the minimum of the number of transmit antennas and the number of receive antennas.

MIMO can offer three kinds of technical benefits:

- *Spatial Multiplexing:* Transmission of multiple data streams simultaneously at the same frequency by means of multiple spatial layers created by multiple antennas. As a result, higher spectral efficiency is obtained.
- *Spatial Diversity:* Use of independent propagation paths enabled by multiple antennas to improve the reliability of the signal transmission against the effect of multiple fading (Gesbert et al., 2003).

- *Array Gain:* Concentrate the transmitted energy of an antenna array in particular directions to improve the received power of desired signals or suppress co-channel interference.

MIMO technologies played a pivotal role in the advancement of 4G radio transmission. To address limitations in hardware, cost, and energy consumption, deploying multiple antennas at the base station became a more appealing and feasible option. Both LTE and WiMAX incorporated multi-antenna technology to implement diverse transmission schemes, e.g.,

- Transmit Diversity: In this MIMO scheme, a single-layer stream is transmitted across multiple signal paths that have low correlation. Spatial diversity is achieved by utilizing either sufficiently large spacing between antennas or employing different antenna polarizations. This scheme is particularly valuable in scenarios that require high reliability, such as control and broadcast channels. In LTE, transmit diversity was specifically defined for either two or four transmit antennas. The transmit symbols can be encoded in a flexible manner using various techniques, such as space-time block coding (STBC) (Alamouti, 1998), space-frequency block coding (SFBC), frequency-switched transmit diversity (FSTD), and cyclic delay diversity (CDD).
- Beamforming: When the downlink channel information is known, the transmitter can employ beamforming in addition to transmit diversity. A narrow beam allows the concentration of transmit energy toward specific directions, resulting in a power gain of the received signal that is proportionate to the number of transmit antennas. Simultaneously, it suppresses interference originating from other directions. Beams can be formed using either high-correlation antenna arrays, known as *classical beamforming*, or low-correlation arrays through the application of *transmit precoding*.
- Open-Loop Spatial Multiplexing: This MIMO scheme used within the LTE system transmits two data streams over two or more antennas. There is no channel information from the UE except a transmit rank indicator (TRI) used to determine the number of spatial layers.
- Closed-Loop Spatial Multiplexing: The UE provides implicit channel state information (CSI) in the form of precoding matrix indicator (PMI), which is fed back to the eNodeB. The feedback information helps in selecting the most suitable precoding matrix. A codebook, which is a collection of predetermined precoding matrices, is known by both the transmitter and receiver. By adjusting the precoded symbols based on the channel conditions, this approach maximizes transmission capacity and allows the receiver to effectively separate different data streams.
- Transmit Antenna Selection: The number of RF chains in a multiple-antenna system generally equals the number of antennas, imposing the challenges of high complexity, high hardware cost, and high power consumption. Given a feedback channel from the receiver to the transmitter. At any time, a single (or a few) antennae with the highest SNR are chosen to transmit the signal. The number of required RF chains is drastically reduced, bringing a significant benefit in terms of hardware cost and size, implementation complexity, and power consumption. Interestingly, it can achieve the full diversity order that equals the number of all transmit antennae participating in the selection, rather than the number of antennae transmitting signals simultaneously (Yu et al., 2018).

7.3.2 Multi-User Multi-Input Multi-Output/MU-MIMO

Utilizing multiple antennas at both the transmitter and receiver, along with precoding and detection processing, serves the purpose of separating spatially multiplexed data streams and suppressing interference among different transmission layers. This technique is commonly referred to as *single-user MIMO* or SU-MIMO (Foschini & Gans, 1998). As an extension of spatial multiplexing, parallel transmission layers created by multiple transmit antennas can be intended for different users with either a single antenna or a few receiver antennas, and vice versa.

In the realm of mobile communication systems or wireless local area networks, the term *multi-user MIMO* or MU-MIMO is used to describe a deployment scenario where a base station or an access point equipped with multiple antennas simultaneously communicates with multiple terminals over the same time-frequency resource. A group of terminals with only one or a few antennas can collaboratively form a virtual array, allowing for the exploitation of spatial multiplexing gains alongside a multi-antenna base station, as illustrated in Fig. 7.3. The base station, with its relatively powerful signal processing capability and sufficient power supply, shoulders the responsibility of spatially separating parallel streams. Therefore, the base station engages in precoding or transmit beamforming for multiple users in the downlink and employs multi-user detection in the uplink (Choi & Murch, 2004).

By breaking up spatial-multiplexed streams among multiple terminals, MU-MIMO gains superiority over SU-MIMO with several fundamental advantages. First, MU-MIMO overcomes a major limitation of SU-MIMO that needs a powerful terminal. The spatial-multiplexing gain is retained even when using low-cost terminals with a limited number of antennas.

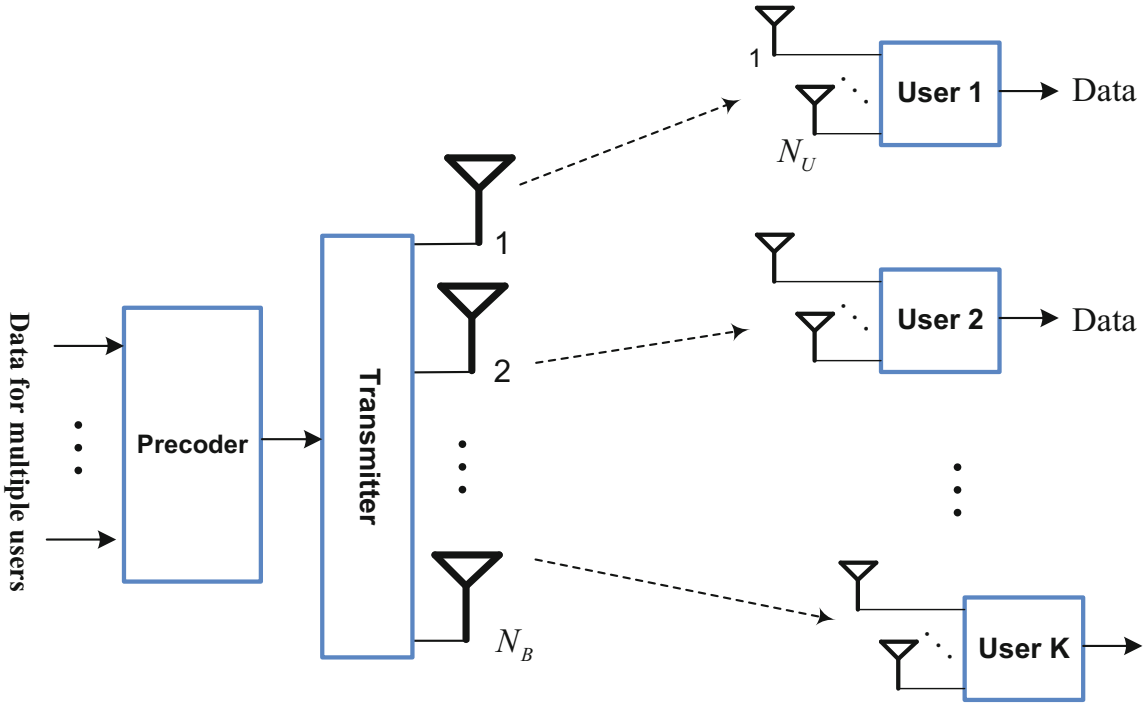


Fig. 7.3 Schematic diagram of an MU-MIMO wireless system, where a BS equipped with N_B antennas simultaneously communicates with K users that have N_U antennas each

This aspect is crucial for achieving economies of scale in the mobile industry. Second, it is less vulnerable to propagation environments due to the spatial distribution of terminals, even under line-of-sight conditions. The achievement of spatial-multiplexing gain is heavily reliant on well-conditioned channels. In a SU-MIMO system, the spatial signatures of the antennas must exhibit decorrelation, which necessitates rich-scattering environments with significant inter-antenna spacing or the utilization of antenna polarization. In an MU-MIMO system, the spatial signatures of different terminals naturally experience decorrelation due to the highly distributed nature of these terminals. As the users are geographically separated, signals propagate in different directions, even in environments with limited scattering. Third, there is an additional benefit of multi-user diversity gain in an MU-MIMO setup, in addition to the spatial-multiplexing gain. Information-theoretic analyses reveal that the sum capacity of multi-user transmission is higher than the channel capacity of single-user communication (Caire & Shamai, 2003).

However, realizing the full potential of MU-MIMO relies on the acquisition of accurate CSI at the transmitter. Research has demonstrated that even a small amount of feedback can greatly enhance the ability to steer power toward the receiver's antennas. Specifically, in SU-MIMO systems, the accuracy of CSI results in a SNR penalty but does not impact the multiplexing gain. However, in an MU-MIMO system, the accuracy of available CSI at the transmitter does affect the multiplexing gain (Spencer et al., 2004). Therefore, it is crucial to provide the transmitter with precise and timely CSI, which poses challenges due to constraints on feedback resources or the severity of wireless channels. The early release of LTE focused on transmit diversity and SU-MIMO. The basic support for MU-MIMO was also provided in Release 8, but it adopted the same codebook-based scheme for SU-MIMO using implicit CSI. The enhanced support for MU-MIMO was added in later releases by the introduction of UE-specific reference signals. Explicit CSI feedback removes the limit of the codebook and provides flexibility to apply advanced transmission schemes such as zero-forcing precoding and conjugate beamforming.

7.3.3 Orthogonal Frequency-Division Multiplexing/OFDM

One of the trends in wireless communications is the increasing widening of the signal bandwidth to support higher transmission rates. In traditional single-carrier transmission, a wider bandwidth in the frequency domain corresponds to

a shorter symbol period in the time domain. With the decrease of the symbol period, the delay spread in a multipath fading channel raises severe ISI and substantially constrains the achievable transmission rate (Jiang & Kaiser, 2016a). To address this issue, a digital filter known as an equalizer is typically used at the receiver to reverse the distortion caused by the channel impairment. The number of filter taps required for the equalizer is proportional to the signal bandwidth. In scenarios where the signal bandwidth is extremely large, an equalizer with hundreds of taps may be needed to effectively mitigate ISI. Unfortunately, implementing such a complex equalizer in practical systems is challenging.

The wireless research community has been seeking an alternative to replace the single-carrier transmission, i.e., multi-carrier modulation (MCM), which is a broadband communication technique that involves dividing a wideband signal into multiple orthogonal narrowband signals. In an MCM system, the symbol period of each narrowband signal is significantly extended and much longer than that of the original wideband signal. Consequently, when the delay spread becomes negligible compared to the extended symbol period, the impact of ISI can be substantially alleviated in an MCM system.

Since its first proposal by Chang (1966), OFDM, or orthogonal frequency-division multiplexing, has been gradually adopted and finally became the dominant modulation technique in wired and wireless communication systems over the past several decades. This is due to its remarkable ability to handle multipath frequency-selective fading without requiring complex equalization, while also offering a straightforward implementation through the use of the discrete Fourier transform (DFT). OFDM has found extensive application in various well-known standards such as Digital Subscriber Line (DSL), Digital Video Broadcasting-Terrestrial (DVB-T), Wi-Fi, WiMAX, LTE, and LTE-Advanced. After thorough comparisons among different modulation techniques, OFDM has also been chosen as a critical component of 5G NR due to its exceptional balance between performance, complexity, scalability, compatibility, and robustness.

As a kind of MCM, the major features of OFDM transmission, which distinguish it from frequency-division multiplexing of multiple narrowband channels, i.e. FDMA, are

- The use of a typically very large number of orthogonal OFDM subcarriers, instead of only a moderate number of non-overlapping carriers in FDMA.
- The use of simple rectangular pulse shaping, and a *sinc*-shaped subcarrier spectrum.
- Tight frequency-domain packing of the OFDM subcarriers with an inter-subcarrier spacing of $\triangle f = 1/T$, where T is the duration of a symbol period.

During the history of OFDM development, there are two milestones, i.e., using DFT as the modulator and the introduction of the cyclic prefix. While a bank of modulator-correlator pairs is generally utilized to illustrate the fundamental principles of OFDM transmission, it is not a practical structure for implementation. In reality, OFDM enables a low-complexity implementation through its specific design and the careful selection of inter-subcarrier spacing. DFT was applied to generate and detect OFDM signals efficiently, which can be further expedited by utilizing a computationally efficient algorithm known as fast Fourier transform (FFT) when the number of subcarriers is a power of two. Furthermore, the concept of cyclic prefix, also known as cyclic extension, was introduced by Peled in his work on resisting time dispersion (Peled & Ruiz, 1980). The insertion of a cyclic prefix involves copying and inserting the last portion of an OFDM symbol at the beginning of the symbol. As long as the span of the delay spread does not exceed the length of the cyclic prefix, the ISI can be absorbed, while the inter-carrier interference can also be avoided since the subcarrier orthogonality is preserved during an integration interval. In addition, the employment of a cyclic prefix can convert the linear convolution with a channel filter into *circular convolution*, also known as *cyclic convolution*, allowing for simple frequency-domain signal processing.

The drawback of cyclic-prefix insertion is the loss of power and bandwidth as the OFDM symbol rate reduces. One way to minimize such a loss is to reduce the inter-subcarrier spacing, with a corresponding increase in the symbol period as a consequence. However, this will increase the sensitivity of the OFDM transmission to fast channel fluctuation due to high Doppler spread. It is also essential to understand that the cyclic prefix does not necessarily have to cover the entire length of the channel time dispersion. In general, there is a trade-off between the power loss and the signal corruption (inter-symbol and inter-subcarrier interference) due to the residual time dispersion not covered by the cyclic prefix.

7.3.4 Orthogonal Frequency-Division Multiple Access/OFDMA

The efficient allocation of radio resources among different active users within a cellular network is crucial due to the limited availability of time-frequency resources. This allocation is a critical aspect in both the uplink and downlink channels, considering that bandwidth is typically scarce and costly. The share of a communications channel among geographically

dispersed multiple users is known as *multiple access*. Common multiple access techniques involve the orthogonal or non-orthogonal partitioning of signaling dimensions into channels, which are then assigned to different users. The multiple access techniques adopted in wireless systems before the development of 4G include TDMA, FDMA, CDMA, and space-division multiple access (SDMA).

The discussion on OFDM has primarily focused on the assumption of point-to-point communication links for the sake of simplicity. In this scenario, all OFDM subcarriers are utilized to multiplex data intended for a single user in the downlink transmission, and in the uplink transmission, a single user is assigned to all subcarriers. However, due to the independence among subcarriers, OFDM transmission can also function as a user-multiplexing or multiple access scheme, enabling simultaneous frequency-separated transmissions involving multiple users. This efficient multiple access scheme is known as OFDMA, which extends the multi-carrier technology of OFDM to provide orthogonal time-frequency resources among multiple users. In the downlink, specific subsets of OFDM subcarriers are dedicated to transmitting data to individual users, as illustrated in Fig. 7.4, while in the uplink, different users can simultaneously transmit their data using separate subsets of OFDM subcarriers.

OFDMA leverages the frequency domain as an additional degree of freedom, resulting in enhanced system flexibility in the following ways:

- It enables the support of scalable system bandwidths without the need to modify fundamental system parameters or equipment design. This significantly improves deployment flexibility in smaller or fragmented frequency bands and allows for seamless expansion of system capacity.
- It allows for the agile allocation of time-frequency resources to different users and implements frequency-domain scheduling to harness the benefits of frequency-diversity gain.
- It facilitates fractional or soft frequency reuse and aids in inter-cell interference coordination.

Just like any technology, OFDMA has its limitations. When OFDMA is utilized as an uplink multiple access scheme, it requires the transmitted signals from different users to reach the base station with minimal time arrival differences. Specifically, the time arrival difference should be within the length of the cyclic prefix to maintain subcarrier orthogonality and prevent inter-carrier interference among users. Considering the variations in propagation delays, it becomes necessary to regulate the uplink transmission timing of each user. For instance, users located farther away from the base station may need to send their signals in advance. To achieve this, transmission-timing control is employed to adjust the transmit timing of each user, ensuring that uplink transmissions are approximately time-aligned at the base station. Moreover, since the propagation

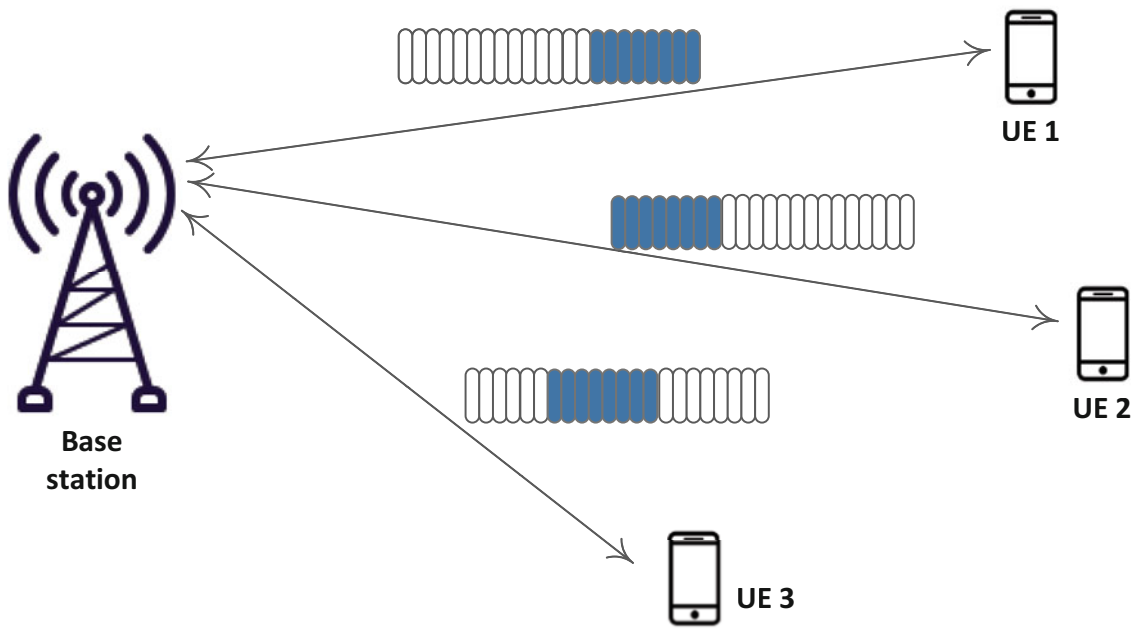


Fig. 7.4 Illustration of an OFDMA-based wireless system, where three spatially distributed users are assigned to three non-overlapping portions of OFDM subcarriers

time changes as users move within the cell, transmission-timing control needs to be a dynamic process. It continuously adapts and fine-tunes the precise timing of each user to account for the changing propagation characteristics.

In addition, even with precise transmission-timing control, inter-carrier interference can still occur due to frequency errors. This interference is generally minimal when frequency errors and Doppler spread are within reasonable limits. However, it assumes that different subcarriers are received at similar power levels. In the uplink, propagation distances and corresponding path losses can vary significantly. As a result, received signal strengths may differ substantially, leading to potential interference from stronger subcarriers to weaker adjacent subcarriers unless perfect subcarrier orthogonality is maintained. To mitigate this interference, some form of uplink power control may be necessary for OFDMA. This involves reducing the transmission power of user terminals located closer to the base station and ensuring that all received signals are approximately at the same power level (Dahlman et al., 2011). By implementing power control, the potential for interference caused by imbalances in received signal strengths can be minimized, thereby preserving subcarrier orthogonality and improving overall system performance.

Following the initial assessment of 3GPP proposals, two potential schemes emerged as candidates for the downlink of the LTE air interface: OFDMA and multi-carrier code-division multiple access (MC-CDMA). On the other hand, for the uplink, the candidate schemes were SC-FDMA, OFDMA, and MC-CDMA. During the 3GPP RAN plenary meeting in December 2005, the selection of multiple access schemes was finalized. OFDMA was chosen as the scheme for the downlink, while SC-FDMA was selected for the uplink. This decision marked an important milestone in the development of 3GPP LTE technology.

7.3.5 Single-Carrier Frequency-Division Multiple Access/SC-FDMA

One of the main technical obstacles in OFDM transmission, like any multi-carrier modulation, is the substantial fluctuations in the instantaneous power of the transmitted signals. These power variations are typically quantified by the peak-to-average power ratio (PAPR). A high PAPR results in reduced efficiency and high cost of the power amplifier. It imposes a specific design constraint for uplink transmission, where low-power consumption and low cost are required for mobile devices. Various techniques have been proposed to address this issue, such as *tone reservation* where certain OFDM subcarriers are not utilized for data transmission and instead modulated to suppress the highest peaks, and *selective scrambling* which selects a transmitted signal with the lowest PAPR from a set of scrambled signals using different codes. However, most of these methods have limitations in their ability to significantly reduce power variations. Therefore, considering a wider-band single-carrier transmission, which maintains a constant envelope with very low PAPR, is an appealing alternative to multi-carrier transmission, particularly in the uplink for mobile terminals.

Figure 7.5 illustrates the fundamental concept at the core of discrete Fourier transform spread OFDM (DFT-s-OFDM) transmission. One way to comprehend DFT-s-OFDM is to perceive it as a variation of conventional OFDM with DFT-based precoding. Similar to OFDM modulation, DFT-s-OFDM operates by generating signals in blocks. In DFT-s-OFDM, a block comprising M modulation symbols from a specific modulation alphabet (such as QPSK or 16QAM) is initially processed through a DFT of size M. Subsequently, the output of the DFT is directed to consecutive subcarriers of an OFDM modulator. In practical terms, the OFDM modulator is typically implemented as an inverse DFT (IDFT) of size N, with any unused inputs of the IDFT set to zero. If the DFT size M equals the IDFT size N, the successive cascaded DFT/IDFT operations would fully cancel each other out. However, when M is smaller than N and the remaining inputs to the IDFT are set to zero, the IDFT output assumes the characteristics of a single-carrier signal. Consequently, the signal exhibits minimal power variations, and its bandwidth is determined by M.

The primary advantage of DFT-s-OFDM over regular OFDM lies in its ability to alleviate sudden fluctuations in transmission power, thereby potentially enhancing power-amplifier efficiency. DFT-s-OFDM is a transmission technique that integrates the favorable characteristics including:

- Minimal fluctuations in the instantaneous power of the transmitted signal, akin to a *single-carrier* property.
- Potential for efficient and high-quality equalization in the frequency domain.
- Capability for flexible bandwidth allocation in an FDMA system.

By dynamically adjusting the block size M, the instantaneous bandwidth of the transmitted signal can be varied, allowing for flexible bandwidth assignment. Furthermore, by assigning the DFT output to a different subset of OFDM subcarriers, the transmitted signal can be shifted in the frequency domain. Multiple users can simultaneously transmit their data by using the DFT-s-OFDM transmission, enabling not only low-power variations like single-carrier transmission but also orthogonal

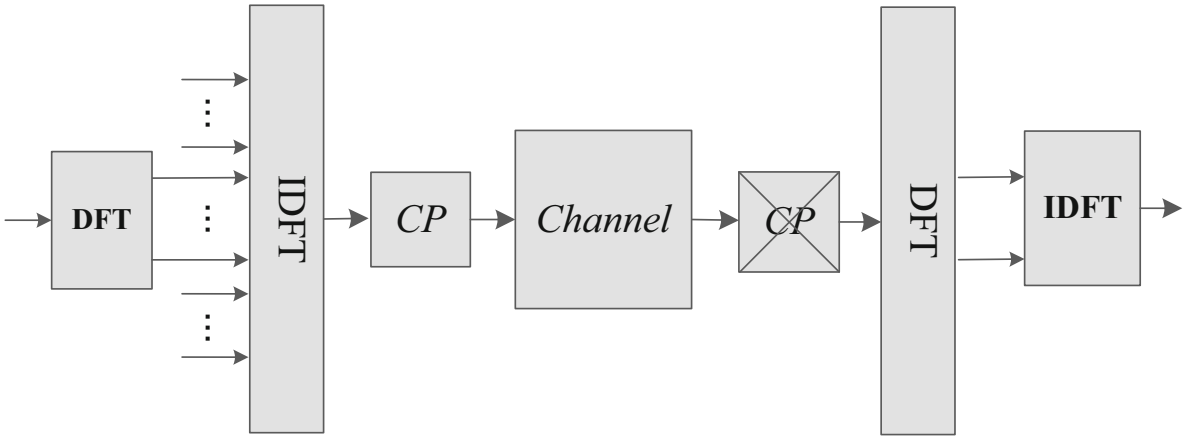


Fig. 7.5 Block diagram of DFT-s-OFDM transmission

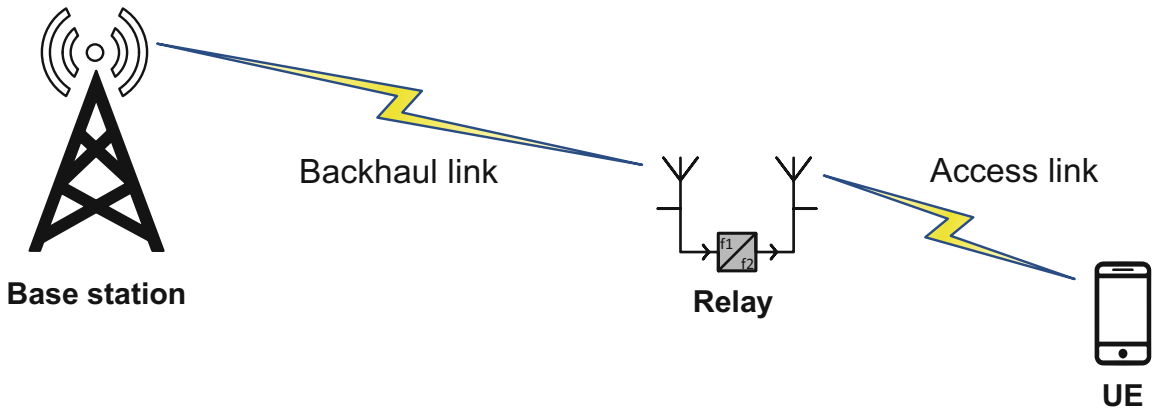


Fig. 7.6 Backhaul and access links in the relaying-based architecture

frequency-division multiple access. Therefore, this technique is referred to as SC-FDMA (Berardinelli et al., 2008), which has been adopted as the uplink transmission scheme in 3GPP LTE.

7.3.6 Relaying

To achieve higher data rates, a terminal needs a relatively high received signal power. The main factor influencing the performance of the connection is path loss, which is greatly affected by the propagation distance. If it is not possible to increase the link budget, such as by using higher transmit power or employing antenna arrays for beamforming, a denser infrastructure is necessary to reduce the propagation distance. Relaying is one effective method that can decrease the distance between the terminal and the infrastructure, leading to an improved link budget. A fundamental requirement for deploying relaying is that the relay nodes should be transparent to the terminals. This means that, from the perspective of the terminal, the relay node behaves like an ordinary low-power base station. This transparency is an important feature as it simplifies terminal implementation and ensures backward compatibility with legacy systems. Additionally, the relay node can access the infrastructure to obtain wireless backhaul, similar to a regular user terminal. As illustrated in Fig. 7.6, the connection between the base station and relay, as well as the relay and terminal, are referred to as the *backhaul link* and *access link*, respectively.

In LTE-Advanced, Release 10 introduced support for *decode-and-forward* relaying, while a simpler scheme known as *amplify-and-forward* relaying does not require additional standardization.

- *Amplify-and-forward* relays, commonly referred to as repeaters, operate by amplifying and forwarding received analog signals. In certain markets, repeaters are commonly used to address coverage gaps. Typically, once deployed, repeaters continuously transmit the received signal, regardless of whether there is a terminal within their coverage area or not. However, more advanced repeaters may also be available. Repeater functionality does not impact the terminal or base station, allowing for their introduction into existing networks seamlessly. It is important to note that repeaters amplify everything they receive, including both useful signals and interference. Consequently, repeaters are most useful in scenarios with high SNR.
- *Decode-and-forward* relays operate by decoding and re-encoding the received signal before forwarding it to the intended users (Jiang et al., 2014). Unlike repeaters, decode-and-forward relays do not amplify noise and interference. This makes them particularly valuable in low-SNR environments. Additionally, independent rate adaptation and scheduling can be implemented for both the access and backhaul links. However, the decode-and-forward relaying has high implementation costs and introduces a larger delay compared to amplify-and-forward repeaters.

Furthermore, relaying is also classified into *inband* and *outband* in terms of the frequency bands used for backhaul and access links.

- In *outband* relaying, the backhaul link operates in a different spectrum from that of the access link, but both use the same radio interface. By ensuring a significant frequency separation between the backhaul and access links, self-interference can be avoided, and the required frequency-domain isolation is achieved. There are no restrictions on the activity of the access and backhaul links, and the relay can potentially operate in full duplex mode.
- In *inband* relaying, both the backhaul and access links operate within the same spectrum. Depending on the relay's deployment and operation, additional mechanisms may be required to prevent interference between the access and backhaul links, as they share the same frequency range. If proper antenna arrangements cannot sufficiently handle this interference, such as when the relay is deployed in a tunnel with the backhaul antenna placed outside the tunnel, a mechanism is needed to separate the activity on the access and backhaul links in the time domain.

7.3.7 Carrier Aggregation

LTE-Advanced and IEEE 802.16 m were developed with the aim of fulfilling the requirements set by IMT-Advanced, which include supporting a maximum bandwidth of at least 40 MHz and a peak data rate of 1 Gbps. However, in most cases, it is difficult to obtain such large portions of the continuous spectrum due to intense competition for spectrum utilization and the fragmentation of legacy spectrum allocation. To overcome this limitation, IMT-Advanced adopted carrier aggregation to enable the implementation of wide bandwidths. For example, LTE-Advanced allows the aggregation of up to five component carriers, which can have different bandwidths, to achieve a maximum transmission bandwidth of 100 MHz for a single terminal (Yuan et al., 2010). Each of these aggregated carriers follows an LTE Release 8 structure, ensuring backward compatibility to support existing UEs. As a result, an LTE terminal can be served transparently using a single-component carrier without any modification, while an LTE-Advanced terminal can achieve higher data rates by utilizing multiple component carriers simultaneously.

The approaches for carrier aggregation can be categorized into three main types:

- *Intra-Band Contiguous*—This type involves using contiguous component carriers within the same frequency band. It is the simplest method and offers advantages such as saving spectral resources that would otherwise be used as guard bands. Additionally, if the RF transceiver has a wide enough bandwidth, a single base-band processing chain can be applied. However, in practice, achieving intra-band contiguous aggregation is often challenging due to the fragmentation of frequency allocation.
- *Intra-Band Non-Contiguous*—When contiguous component carriers are not available, LTE-Advanced supports intra-band non-contiguous carrier aggregation. It allows for the utilization of fragmented frequency bands. In this case, the component carriers belong to the same band but have a frequency gap or gaps in between. This method enables the efficient use of available spectrum resources, even if they are not contiguous within the band.
- *Inter-Band*—With inter-band non-contiguous carrier aggregation, the component carriers are not only fragmented but also belong to different operating frequency bands. It offers the advantage of frequency diversity, as different frequency bands correspond to different channel fading characteristics. However, achieving inter-band non-contiguous aggregation requires the use of several independent RF and base-band processing chains, which increase the complexity and costs of the implementation.

These different approaches for carrier aggregation provide flexibility in utilizing spectrum resources and enable the deployment of wider bandwidths, thereby enhancing the capacity and performance of IMT-Advanced networks. The choice of aggregation approach depends on factors such as the availability of contiguous spectrum, frequency band fragmentation, and the desired trade-offs between complexity, bandwidth utilization, and frequency diversity.

7.3.8 Coordinated Multi-Point/CoMP Transmission and Reception

Mobile users have high expectations for receiving excellent QoS regardless of their location or the time. In a cellular network, a BS is placed at the center of a cell of a network of cells and serves a lot of users simultaneously. High QoS is primarily offered to users close to the BS, namely the cell center. The area of a cell center is only a small portion of the whole coverage. Many users are at the cell edge, suffering from worse QoS due to weak signal strength, which is restricted by large distance-dependent path loss, strong inter-cell interference, and handover issues that are inherent to the cellular architecture. The performance gap between the center and edge in a cell is not trivial; however, it is tremendous. For example, ITU-R M.2410 (2017) specified the minimum requirements related to technical performance for IMT-2020, where peak spectral efficiency in the downlink and uplink reaches 30 bps/Hz and 15 bps/Hz, respectively. The so-called 5th percentile spectral efficiency, a.k.a 95%-likely spectral efficiency, can be guaranteed to 95% of the users and thus define user-experienced QoS at the cell edge. In contrast, the targets of 5th percentile spectral efficiency are only 0.3 bps/Hz (downlink) and 0.21 bps/Hz (uplink) in indoor hotspot eMBB, which are further lowered to 0.12 bps/Hz and 0.0453 bps/Hz in rural eMBB, amounting to a huge difference more than 100 times (Jiang & Schotten, 2023c).

To improve cell-edge performance, various techniques for mitigating inter-cell interference have been proposed over the years, including interference averaging, frequency hopping, and interference coordination. One notable example is inter-cell interference coordination (ICIC), along with its enhanced version called eICIC, which was introduced by LTE-Advanced as a technical feature. Another approach considered by 3GPP to enhance the spectral efficiency at the cell edge is CoMP transmission and reception.

CoMP can be regarded as an extension of ICIC techniques aimed at achieving faster decision-making and wider backhaul bandwidth. These improvements facilitate the sharing of real-time CSI and data among transmission points. In 3GPP studies, CoMP techniques are typically categorized into three types based on the constraints imposed on the backhaul link between coordinated points and the complexity of scheduling. The coordination between multiple geographically separated sites can be performed in either the downlink or the uplink, resulting in varying levels of coordination. These categories include coordinated scheduling, also known as coordinated beamforming, joint processing, and dynamic point selection (Irmer et al., 2011).

- *Joint Processing:* Within the coordination area, multiple sites engage in simultaneous transmission or reception of signals to and from a single UE. This approach enhances the quality of the received signal, mitigates or prevents inter-cell interference, and achieves macro-diversity gain. However, implementing this method necessitates a robust backhaul network due to the necessity of exchanging transmitted or received data, channel information, and calculated transmission weights among the coordinated sites.
- *Dynamic Point Selection:* It is a technique where UE data is accessible to multiple base stations within a coordinating set, but only one base station transmits or receives the signals at a given time/frequency. The selected base station is referred to as the transmission point and can be changed from one subframe to another to ensure optimal transmission for a UE with fluctuating channel conditions. This situation commonly occurs at the cell edge, where the serving base station benefits from long-term channel characteristics, while other cooperating base stations may have more favorable short-term characteristics.
- *Coordinated Scheduling/Beamforming:* A coordinated set is formed by base stations collaborating to enhance system performance, particularly at the cell edges. These base stations engage in the exchange of information regarding UE and their respective channel conditions. The shared information is utilized to make coordinated scheduling decisions, with the collective goal of improving overall system capacity and providing enhanced QoS for UEs. Conversely, coordinated beamforming focuses on optimizing signal transmission from multiple base stations to a specific UE. This technique involves aligning the transmission beams of the base stations toward the intended UE through the application of advanced beamforming algorithms. By coordinating their beamforming strategies, the base stations can maximize the signal strength at the UE and minimize interference from neighboring cells, resulting in an improvement in the received signal quality of the UE and enhancing the overall performance of the system.

7.3.9 Low Density Parity-Check/LDPC Codes

In 1948, the foundation of information theory was laid by Claude E. Shannon through the publication of his well-known paper (Shannon, 1948). Shannon's groundbreaking work explored the constraints of reliably transmitting data over unreliable channels. He proved the existence of a capacity, a numerical threshold specific to each communication channel. Transmission rates arbitrarily closing to this capacity allow for reliable data transfer, while rates above capacity make reliable transmission impossible. Shannon also demonstrated the presence of capacity-achieving codes, where the probability of error for the maximum likelihood decoder approaches zero as the code's block length goes to infinity. The crucial role of developing efficient encoding and decoding algorithms for these codes, striving to approach the channel's capacity, significantly influenced the evolution of wireless communications.

In his 1960 Ph.D. dissertation, Robert Gallager invented Low-Density Parity-Check codes (Gallager, 1962). Despite their potential, these codes initially went unnoticed due to the challenges posed by high-complexity computation. The concatenated Reed–Solomon and convolutional codes garnered recognition as highly effective for error control coding, diverting attention from LDPC codes. However, this technique experienced a resurgence after three decades. This intriguing aspect of their history unfolded as two distinct communities, see MacKay and Neal (1997) and Richardson et al. (2001), independently reinvented analogous LDPC codes around the same period, driven by entirely different motivations.

The key characteristic of LDPC codes is the sparse parity-check matrix, allowing for efficient encoding and decoding algorithms. The encoding process involves multiplying the information bits by the sparse parity-check matrix to produce the codeword. The encoding begins with a vector of information bits, known as the message vector. The sparse parity-check matrix, characterized by a low density of 1s, defines relationships between variable nodes and check nodes. Variable nodes represent the message bits, while check nodes embody parity-check equations. The encoding involves multiplying the message vector by the sparse matrix, using modulo-2 addition (XOR) to calculate parity-check values. The resulting codeword, comprising both original information bits and parity-check bits, is systematically generated. Iterative algorithms, such as the belief propagation algorithm or the sum-product algorithm, are commonly employed in the decoding process. The primary objective is to iteratively enhance the accuracy of transmitted bit probabilities and rectify errors introduced by the communication channel. The initialization phase sets probabilities for each bit (0 or 1), and subsequent iterations involve dynamically updating these probabilities through message exchanges between variable nodes and check nodes. The parity-check matrix's sparse nature facilitates efficient computation, enabling the iterative refinement of estimates. The decoding process persists until a specified stopping criterion is met, such as reaching a maximum iteration limit or attaining a predefined convergence level.

LDPC codes have demonstrated excellent performance. An irregular LDPC code with a length of one million, constructed by Richardson et al. (2001), achieved a bit-error rate of 10^{-6}, less than 0.13, dB away from the best possible as determined by the Shannon capacity. This surpasses the performance of previously best-known Turbo codes. Moreover, Chung et al. (2001) developed another LDPC code that achieved results within 0.04 dB of the Shannon limit at a bit-error rate of 10^{-6}, using a block length of 10^7.

7.3.10 Heterogeneous Network/HetNet

The network architecture has been transformed due to the increasing need to deliver high data rates to a large number of mobile users. In cellular networks, to achieve high system capacity, various techniques are employed. These include employing advanced radio transmission methods to enhance spectral efficiency, acquiring more spectral resources, or deploying denser network nodes. In a traditional homogeneous network, cell splitting is used to decrease cell size and increase capacity, but this requires network re-planning and system reconfiguration. Additionally, the traffic demand is not evenly distributed. To address these challenges, LTE-Advanced introduced Heterogeneous Networks (HetNets), which consist of cells of different sizes, transmit powers, coverage areas, and hardware capabilities, such as macro, micro, pico, and femtocells. Low-power, cost-effective base stations are added to macro-cell networks as underlying nodes to provide high capacity in areas with high demand or to fill coverage gaps, both indoors and outdoors.

Ideally, if there are sufficient radio resources available, different frequency bands can be allocated to different cell types to avoid interference. However, due to the increased traffic demand, it is not always possible to allocate a dedicated carrier for small cells. As a result, heterogeneous cells must coexist and operate on the same set of frequencies, even if they overlap geographically. One of the major challenges in deploying a heterogeneous network is managing inter-cell interference. To tackle this issue, 3GPP established a work item focused on ICIC, which was later evolved to eICIC. LTE-Advanced utilizes

carrier aggregation to support the deployment of such a heterogeneous network. In this approach, cross-carrier scheduling is employed, allowing control signaling to be transmitted on a shared component carrier. This helps prevent interference on control channels between the macro-cell layer and the small-cell layer.

7.3.11 Device-to-Device/D2D Communications

In a traditional mobile network, the information is first sent to a base station in the uplink and then forwarded to the target terminal in the downlink, even if two communicating parties are close to each other. As we know, the performance of the transmission link heavily depends on the propagation distance; namely, a shorter distance gains a high SNR. The emergence of new application scenarios such as content distribution and Proximity-based Service (ProSe) defined in 3GPP Release 12 fostered the introduction of D2D communications in cellular networks (Liu et al., 2015). D2D communication establishes a direct link between devices, enabling them to exchange data, share resources, and communicate within close proximity without relying on centralized network infrastructure. This type of direct communication offers numerous advantages, such as high spectral efficiency, system capacity, energy efficiency, low latency, and fairness. D2D communication can be implemented in various wireless networks, including cellular networks, wireless local area networks, and peer-to-peer networks.

The technical advantages of D2D communications include:

- *Enhanced efficiency*: By enabling direct communication between nearby devices, D2D reduces the need to transmit data through base stations or the core network. A direct link reduces latency, improves throughput, and optimizes network resource utilization.
- *Localized communication*: D2D allows devices in close proximity to communicate directly, which is particularly useful in scenarios where devices need to share information or collaborate within a specific area. For example, in crowded events, devices can directly exchange multimedia content or share files.
- *Offloading network traffic*: D2D communication facilitates direct communication between devices, reducing the reliance on the cellular network for data transmission. This offloads network traffic, alleviates congestion, and enhances overall network capacity.
- *Proximity-based services*: Leveraging D2D communication, proximity-based services and applications can be developed. Examples include location-based advertising, multiplayer gaming, file sharing, and collaborative work applications, which enhance user experiences by leveraging the direct communication capabilities of nearby devices.

D2D communication is usually not transparent to the cellular network, and it can operate on the licensed spectrum (i.e., inband) or license-exempt spectrum (i.e., outband).

- *Inband D2D communications*: Both D2D direct and cellular links make use of the same frequency bands. Inband D2D communication can be further classified into two types, namely underlay and overlay, based on the frequencies employed in D2D and cellular links. The underlay type aims to enhance the spectral efficiency of a cellular network by reusing the same spectrum for both types of links. On the other hand, the overlay type necessitates dedicated spectral resources specifically allocated for D2D links, establishing direct connections between transmitters and receivers. However, a significant drawback of inband D2D communications is the interference that may occur between D2D users and cellular communications, causing disruptions in both systems.
- *Outband D2D communications*: The purpose of this approach is to leverage the unlicensed spectrum and mitigate interference between D2D and cellular links. To utilize the unlicensed spectrum, a distinct interface is employed, often employing wireless technologies like Wi-Fi and Bluetooth. While using the unlicensed spectrum helps avoid interference within the band, there is a potential drawback of uncontrolled interference in the unlicensed spectrum itself.

The implementation of D2D communications presents various challenges that must be overcome. Managing interference is crucial since many D2D communications operate in the same frequency bands as the cellular network, necessitating coordination and interference management between D2D and cellular transmissions. Efficient resource allocation is also a concern to prevent conflicts with cellular operations while ensuring fairness and maintaining QoS. Power control algorithms need to be optimized for transmission power levels and interference management. Security and privacy are important considerations to prevent unauthorized access and malicious activities in direct links. Scalability becomes a challenge as the number of D2D-enabled devices increases, requiring scalable algorithms and protocols. QoS provisioning in dynamic and interference-prone environments requires effective scheduling, resource management, and interference mitigation

techniques. Additionally, network overhead may be introduced by D2D communication, impacting network capacity and resource utilization through signaling and control messages. Overcoming these challenges necessitates the development of advanced algorithms, protocols, and coordination mechanisms that optimize D2D communication within the cellular network framework, ensuring efficient resource utilization, effective interference management, and robust security and privacy measures.

7.3.12 License-Assisted Access/LAA

License-assisted access is a technology implemented in LTE-Advanced Pro (Release 13) that enables cellular networks to utilize unlicensed spectrum, including the 5 GHz frequency band, alongside their licensed spectrum. By incorporating unlicensed spectrum, LAA allows mobile network operators to enlarge network capacity and improve overall system performance. In the context of LAA, the primary operations of the cellular network occur within the licensed spectrum, which provides a controlled and regulated environment. However, when additional capacity is required, LAA allows the network to offload data traffic to the unlicensed spectrum.

The underlying principle of LAA is to ensure equitable coexistence and minimize interference between users of licensed and unlicensed spectra. To achieve this, LAA incorporates mechanisms such as channel sensing and the Listen-before-Talk (LBT) protocol, which are similar to those employed in Wi-Fi networks. These protocols enable the network to assess the availability of the unlicensed spectrum and determine the appropriate times for cellular communications to utilize it. Additionally, LAA employs carrier aggregation, wherein a secondary carrier in the unlicensed spectrum, specifically the 5 GHz band, is utilized to offload primary carrier traffic in the licensed band. This combination of licensed and unlicensed bands allows consumers to achieve higher peak data rates both indoors and outdoors, improving the user experience (Thalanany et al., 2017).

7.3.13 Self-Organizing Networks/SON

The deployment of 4G networks faced extra difficulties in operation and management since the coexistence of 4G with 2G and 3G networks resulted in a more complex network structure, with numerous parameters that require configuration and optimization. The increased complexity imposed a significant challenge on traditional manual and semi-automatic network management methods, which were already expensive and time-consuming. At that time, network issues such as equipment configuration, system failures, and network optimization still required manual reconfiguration of software, hardware repairs, or the installation of new equipment. Mobile operators must maintain a sizable operational team of network administrators, leading to high operational expenditure (OPEX) that was three times greater than capital expenditure (CAPEX) and continually rising. Furthermore, troubleshooting cannot be carried out without interrupting network operations, which violates service level agreements and negatively impacts the user's quality of experience (Jiang et al., 2017a).

Self-organizing networks (SON) encompass a set of functionalities that facilitate the automatic configuration, optimization, diagnostics, and recovery of cellular networks. The adoption of the SON paradigm enables the automation of these tasks, empowering the network to adapt seamlessly and dynamically to various scenarios. This approach is regarded as essential when mobile networks and operations become increasingly complex, primarily driven by the mounting cost pressures faced by the industry. The functionalities of initial SON technologies are mainly concentrated into three aspects:

- *Self-Configuration:* During the installation phase, the base station is not yet active, and the signal transmission has not started. In this stage, a new base station can be seamlessly integrated into the network through automatic software installation and database downloads. The objective of self-configuration is to simplify and expedite the integration process of the new node into the network, minimizing the need for manual intervention, especially during on-site commissioning.
- *Self-Optimization*: Self-optimization aims to maximize network performance and achieve greater energy and cost savings compared to manual optimization methods. Through this functionality, the network is automatically fine-tuned using measurements and performance data gathered from user terminals and base stations. The goal is to continuously optimize the network performance and ensure the user's quality of experience without requiring manual intervention.
- *Self-Healing*: Parallel to self-optimization, self-healing functions by autonomously identifying and pinpointing faults, such as issues with sites and cells, and rectifying any losses in coverage and capacity. Through modifying parameters

and algorithms in neighboring cells, self-healing strives to reinstate functionality in faulty cells. It integrates automated troubleshooting with subsequent network self-optimization to uphold a resilient and dependable network performance.

SON technologies were introduced in 3GPP Release 9 with the purpose of effectively reducing both CAPEX and OPEX in the LTE system. The initial SON features assisted mobile operators in deploying LTE networks by clustering eNodeBs within existing 2G and 3G legacy networks, enabling them to meet the initial coverage requirements. As LTE networks expanded and achieved wider coverage, mobile operators shifted their focus toward network growth and optimizing capacity and coverage in a heterogeneous environment. The heterogeneous network includes macros, micros, picos, and femtos cells, where 2G and 3G coexist alongside multiple carriers. To further improve the performance of heterogeneous networks and continuously reduce OPEX, additional features are being standardized in 3GPP Release 10 and later versions. These standardized features created additional opportunities to enhance the performance and efficiency of managing heterogeneous networks.

3GPP has spent a lot of effort to establish a range of SON use cases to facilitate lowering the OPEX of LTE networks (Jorguseski et al., 2014). These use cases can be categorized as either distributed or centralized. Centralized SON operates with slower update rates and relies on long-term statistics collected from multiple cells involved in the optimization process. Typically, these centralized SON functions are implemented in operational and management sub-systems, depending on the need for multi-vendor capabilities. In contrast, distributed SON is particularly suitable for processes that require quick response times and have an impact on only a few cells. It is also utilized in scenarios where parameter changes have a localized effect, but information regarding the configuration or status of neighboring cells is required. In distributed SON, information exchange between cells can be facilitated through inter-cell interfaces. Optimization algorithms are executed within the cell in the case of distributed SON. In practical deployments, a hybrid SON approach is often employed, where centralized and distributed SON are integrated simultaneously to accommodate diverse requirements across various use cases. In hybrid SON, some optimization algorithms are executed within the operations and maintenance sub-system, while others are executed directly within the cell.

7.4 Summary

For the first time in the history of the mobile industry, a unified global cellular standard emerged, rather than several competing ones, with the extensive adoption of Long-Term Evolution (LTE) as the *de facto* 4G technology. This happened despite the approval of another IMT-Advanced standard by ITU-R, referred to as WirelessMAN-Advanced, which was evolved from WiMAX, one of the IMT-2000 standards. This chapter aims to review these two IMT-Advanced standards and shed light on the driving forces behind the transition from 3G data-centric cellular networks to 4G mobile broadband. The fundamental technologies empowering the success of 4G, including multiple-input multiple-output (MIMO), orthogonal frequency-division multiplexing (OFDM), orthogonal frequency-division multiple access (OFDMA), single-carrier frequency-division multiple access (SC-FDMA), relaying, and device-to-device (D2D), are elaborated to provide readers with an insightful view.

7.5 Exercises

1. The early generations of cellular systems were designed to offer mobile voice communication services. With the surge of data services, a packet-switched network for data delivery and a circuit-switched network for voice calls were maintained in both the 2.5G and 3G systems. What is the big leap made by 4G?
2. In general, we recognize LTE as a 4G standard. The ITU-R Working Party 5D (WP5D) released the essential technical requirements for the 4G system in 2008, known as the IMT-Advanced, in its recommendation M.2134. Is LTE one of IMT-Advanced standards?
3. What are the multiplexing/multiple access technologies adopted by LTE?
   (A) Frequency-division multiple access (FDMA)
   (B) Time-division multiple access (TDMA)
   (C) Code-division multiple access (CDMA)
   (D) Orthogonal frequency-division multiple access (OFDMA)
   (E) Single-carrier frequency-division multiple access (SC-FDMA)

4. Identify the major technical advantage of SC-FDMA over OFDMA. Explain why it is more suitable for uplink transmission.
5. LTE supports a maximal bandwidth of 20 MHz. To avoid mutual interference between adjacent frequency bands, a pair of guard bands 2×1 MHz is assigned. That means the transmission of OFDM signal occupies 18 MHz. The inter-subcarrier spacing is set to 15 kHz while different DFT sizes, with values being 128, 256, 512, 1024, and 2048 are used. Making a simple calculation, we know that the number of OFDM subcarriers equals

$$\frac{18\,\text{MHz}}{15\,\text{kHz}} = 1200. \tag{7.1}$$

   From the opposite side, another calculation is

$$15\,\text{kHz} \times 2048 = 30.72\,\text{MHz}. \tag{7.2}$$

   Can you explain the mismatch between the DFT size and the number of OFDM subcarriers?
6. One of the symbolic features of 4G was the extensive application of MIMO technology. How many antennas are supported by LTE-advanced?
7. *Amplify-and-forward* relays, commonly referred to as repeaters, operate by amplifying and forwarding received analog signals. *Decode-and-forward* relays operate by decoding and re-encoding the received signal before forwarding it to the intended users. Unlike repeaters, decode-and-forward relays do not amplify noise and interference. In LTE-Advanced, Release 10 introduced the support for *decode-and-forward* relaying. Find the reason why *amplify-and-forward* has not been standardized there.
8. What are the main objectives of applying CoMP transmission and reception?
9. Why LTE-advanced adopted carrier aggregation? What is the maximal bandwidth supported by LTE-advanced?

Long-Term Evolution Advanced (LTE-A)

8

As introduced earlier in Chaps. 1 and 7, 3GPP Release 10 is the first standard that fully meets the 4G requirements of IMT-Advanced. In this chapter we will focus on Release 10 to introduce the key features and designs of LTE-A. Advanced technologies that were first introduced in later releases of LTE-A, such as CoMP, D2D communication, LAA, etc., will not be covered.

8.1 Frequency Bands and Key Features

Similar to UMTS, LTE and LTE-A also operate in various frequency bands, allowing for global deployment and tailored network configurations. The ITU has defined specific frequency bands for Evolved Universal Terrestrial Radio Access Network (E-UTRAN), which is the RAN of LTE/LTE-A. These bands are allocated to different regions worldwide, to provide flexibility for network operators to optimize their deployments based on local requirements. A detailed list of the E-UTRA operation bands is provided in Table 8.1. Like UTRA (see Chap. 6), E-UTRA bands are also clustered into the paired bands for operation in FDD, and unpaired ones for TDD. Since FDD-LTE and TDD-LTE distinguish from each other barely more than the radio frame design, in this chapter we will be mainly focusing on the FDD mode, which is globally more deployed. A summary to the TDD mode will be given in Sect. 8.8.

By utilizing a wide range of frequency bands and incorporating advanced technologies such as carrier aggregation, LTE-A offers a series of new features to enable operators to deliver high speed data services, support multimedia applications, and provide a superior user experience. The key features of LTE-A are including:

- **High data rates**: LTE/LTE-A offers substantially higher data rates compared to previous generations of cellular networks. With peak data rates reaching up to several hundred Mbps in LTE and Gbps in LTE-A, users can enjoy faster downloads, seamless video streaming, and real-time gaming experiences.
- **Enhanced spectral efficiency**: LTE/LTE-A employs advanced modulation techniques, such as OFDM and MIMO, to achieve higher spectral efficiency. This translates to more efficient utilization of available frequency resources, allowing for increased network capacity and improved overall performance.
- **Low latency**: LTE/LTE-A significantly reduces network latency, resulting in faster response times for interactive applications. This low latency is crucial for real-time services like online gaming, video calling, and IoT applications, enabling smooth and responsive user experiences.
- **All-IP network architecture**: LTE/LTE-A adopts a fully packet-switched, all-IP network architecture, eliminating the need for separate circuit-switched networks. This IP-based architecture provides a more efficient and flexible network infrastructure, enabling seamless integration with other IP-based systems and services.
- **QoS differentiation**: LTE/LTE-A incorporates robust QoS mechanisms, allowing operators to prioritize traffic based on specific service requirements. This differentiation ensures that critical applications, such as voice and video, receive the necessary bandwidth and network resources, guaranteeing optimal performance.
- **Multiband operation**: LTE/LTE-A supports operation across multiple frequency bands, allowing for global compatibility and flexible network deployments. By utilizing a variety of frequency bands, operators can optimize coverage, capacity, and performance based on regional spectrum availability and specific deployment scenarios.

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5_8

Table 8.1 E-UTRA operating bands (3GPP TS 36.101 2022)

Operating band	UL frequencies (MHz)	DL frequencies (MHz)	Duplex mode
1	1920–1980	2110–2170	FDD
2	1850–1910	1930–1990	FDD
3	1710–1785	1805–1880	FDD
4	1710–1755	2110–2155	FDD
5	824–849	869–894	FDD
6[a]	830–840	875–885	FDD
7	2500–2570	2620–2690	FDD
8	880–915	925–960	FDD
9	1749.9–1784.9	1844.9–1879.9	FDD
10	1710–1770	2110–2170	FDD
11	1427.9–1447.9	1475.9–1495.9	FDD
12	699–716	729–746	FDD
13	777–787	746–756	FDD
14	788–798	758–768	FDD
15	Reserved	Reserved	FDD
16	Reserved	Reserved	FDD
17	704–716	734–746	FDD
18	815–830	860–875	FDD
19	830–845	875–890	FDD
20	832–862	791–821	FDD
21	1447.9–1462.9	1495.9–1510.9	FDD
22	3410–3490	3510–3590	FDD
23	2000–2020	2180–2200	FDD
24[b]	1626.5–1660.5	1525–1559	FDD
25	1850–1915	1930–1995	FDD
…			
33	1900–1920	1900–1920	TDD
34	2010–2025	2010–2025	TDD
35	1850–1910	1850–1910	TDD
36	1930–1990	1930–1990	TDD
37	1910–1930	1910–1930	TDD
38	2570–2620	2570–2620	TDD
39	1880–1920	1880–1920	TDD
40	2300–2400	2300–2400	TDD
41	2496–2690	2496–2690	TDD
42	3400–3600	3400–3600	TDD
43	3600–3800	3600–3800	TDD

[a] Band 6 is not applicable
[b] DL operation in this band is restricted to 1526–1536 MHz, UL restricted to 1627.5–1637.5 MHz and 1646.5–1656.5 MHz

- **Carrier aggregation**: LTE-A introduces CA, which allows it to bundle multiple carriers, a.k.a. component carriers (CCs), into a single logical channel. Each component carrier operates on a specific frequency band and has its own bandwidth. The aggregated bandwidth is the sum of the individual component carriers, providing a wider transmission channel for data. Carrier aggregation increases available bandwidth, enhances data rates by combining resources from different bands, enables efficient spectrum utilization, and improves network efficiency.
- **Enhanced mobility**: LTE/LTE-A provides seamless mobility support, allowing users to maintain connectivity while moving across cells and different network environments. Advanced handover mechanisms ensure smooth transitions between cells, minimizing call drops and data interruptions during mobility events.

8.2 LTE-A Channels

The logical, transport, and PHYsical channels in LTE/LTE-A systems are illustrated in Fig. 8.1. Compared to the UMTS (see Fig. 6.2), LTE-A has a much conciser channel design.

For user data, LTE/LTE-A canceled the design of CTCH, FACH, and SCCPCH, which were used for low-data rate RRC connection. On the transport layer, the Uplink/Downlink Shared Channels (UL/DL-SCH) are used to provide user data transfer over DCCH and DTCH, instead of the mixture of DCH, HS-DSCH, and E-DCH in UMTS. The PHYsical layer channels are also correspondingly pruned. In addition, multicast/broadcast traffic, which was not an essential part of the general UMTS channels but delivered over the MBMS framework, is in handled in LTE-A by the Multicast Traffic Channel (MTCH), Multicast Channel (MCH), and Physical Multicast Channel (PMCH) on the logical, transport, and PHYsical layers, respectively.

For control signaling, three PHYsical channels are set in the downlink, namely Physical Control Format Indicator Channel (PCFICH), Physical Downlink Control Channel (PDCCH), and Physical Hybrid ARQ Indicator Channel (PHICH), respectively. PCFICH indicates the number of OFDM symbols used for the PDCCH that carries the Downlink Control Information (DCI), and PHICH is used to report the HARQ status. In the uplink, the Physical Uplink Control Channel (PUCCH) provides various control signaling, including the Uplink Control Information (UCI), scheduling request, downlink data ACK/NACK, and CQI. In addition, the Physical Uplink Shared Channel (PUSCH) may sometimes also be used to carry uplink control information.

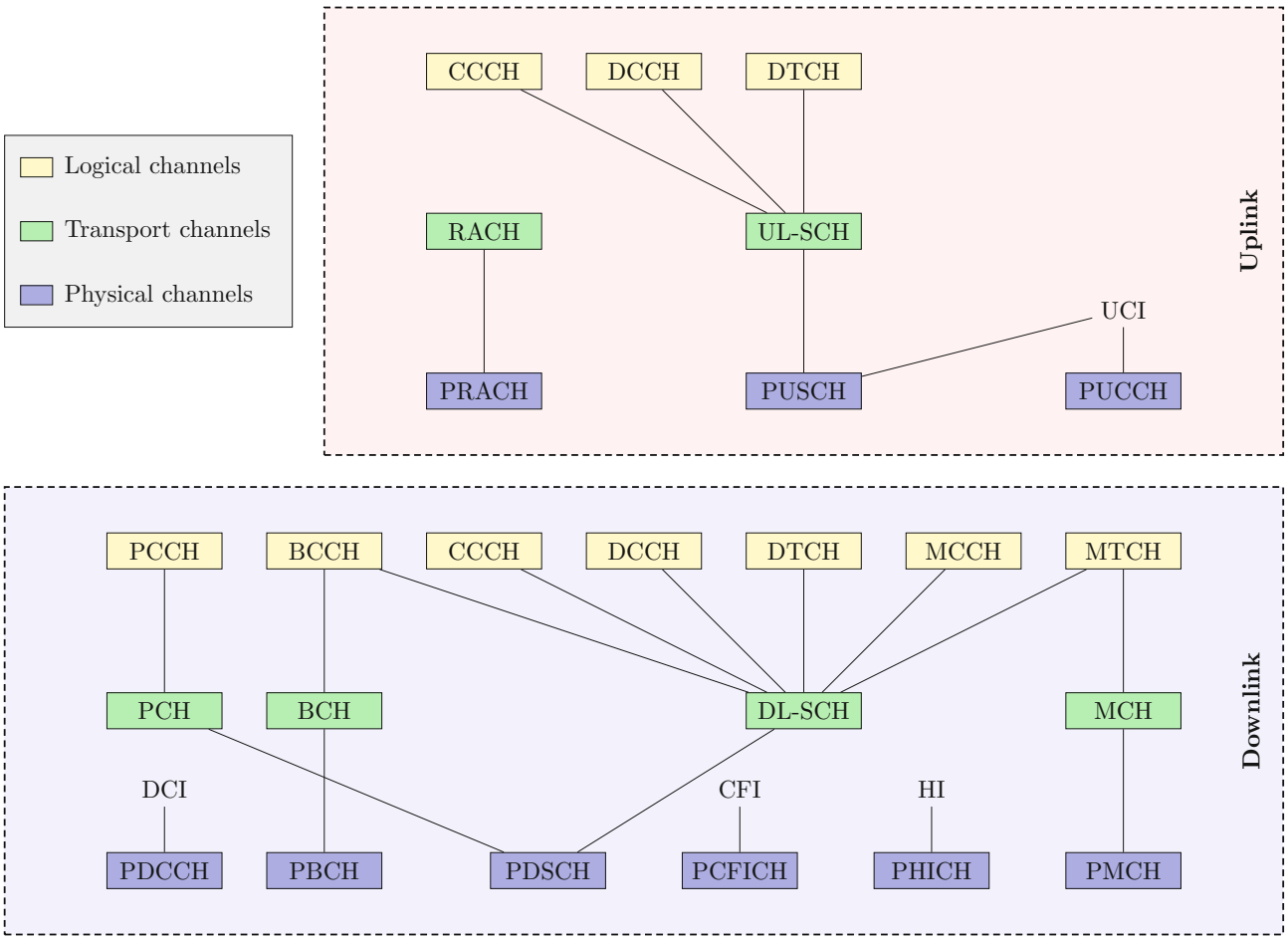


Fig. 8.1 LTE-A channels

8.3 LTE-A Architecture

Like its predecessors, the LTE-A network also consists of the UE, the RAN, and the CN. The CN of LTE/LTE-A is called the EPC, which is designed not only to work with the E-UTRAN but also compatible with the legacy RAN such as UTRAN or GERAN, as shown in Fig. 8.2.

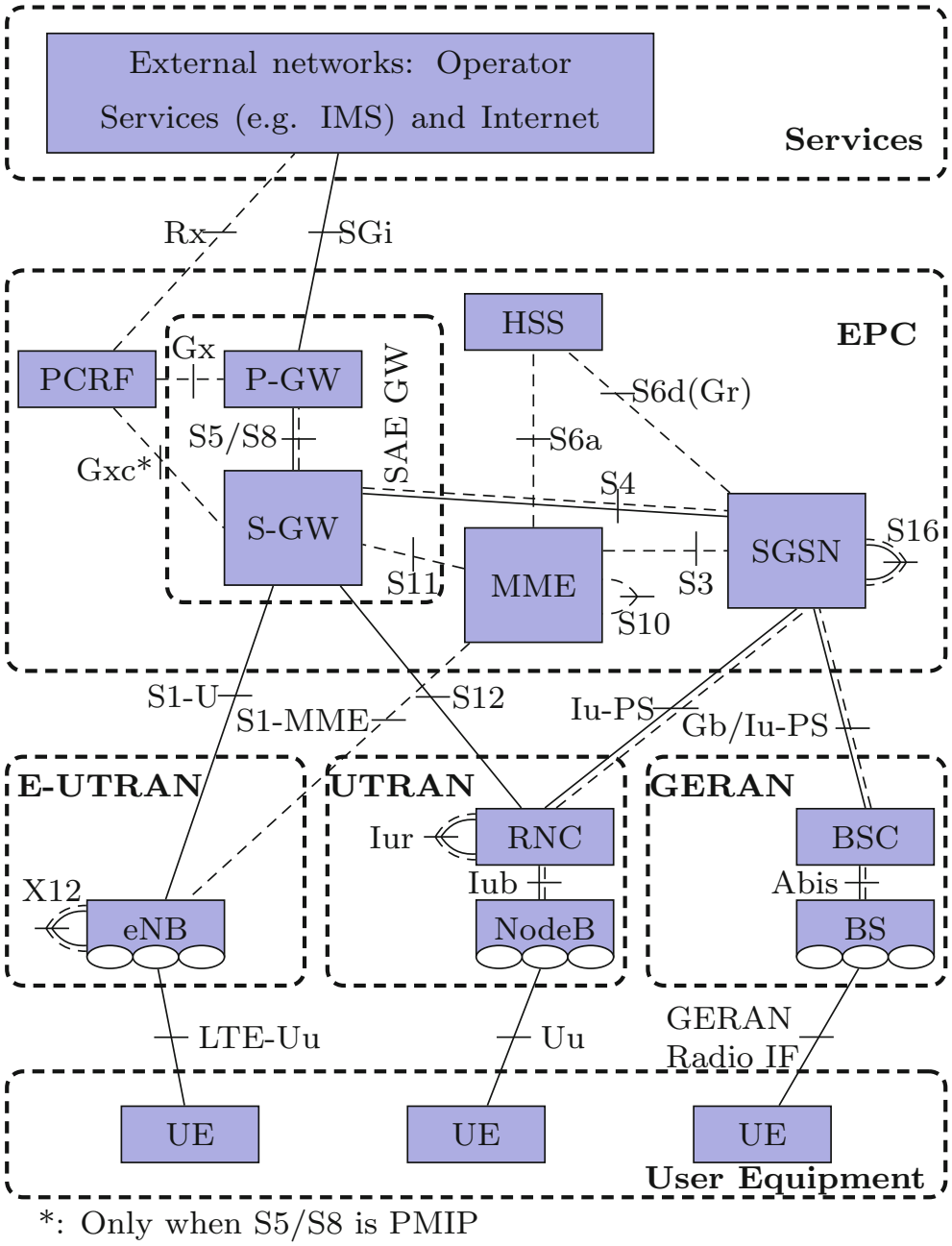


Fig. 8.2 Overall LTE-A architecture

8.3.1 E-UTRAN Node B

The most significant evolution of E-UTRAN is its flat architecture. Unlike UTRAN or GERAN, there is no standalone controller entity like RNC or BSC in the E-UTRAN. Instead, the only entity in the E-UTRAN is the E-UTRAN Node B, or eNB (eNodeB), which takes on the responsibility of both the base station and the control functions previously handled by the RNC.

The eNodeB is equipped with enhanced capabilities to manage radio resources, perform radio resource control, and handle mobility and scheduling functions. It performs several critical functions to ensure efficient operation and management of the radio access network. These functions include:

- **RRM**: The eNodeB is responsible for managing radio resources within its coverage area. It allocates and schedules radio channels, assigns transmission parameters, and optimizes resource utilization to maximize network capacity and performance. RRM functions include power control, admission control, handover control, and interference management.
- **Mobility Management**: The eNodeB handles mobility-related functions such as tracking the location of UEs, initiating handovers between cells, and managing cell reselection procedures. It ensures seamless mobility for UEs as they move across different coverage areas while maintaining the quality of the connection.
- **RLC**: The eNodeB performs RLC functions, which include segmentation and reassembly of data packets, error correction, and flow control. RLC ensures reliable and efficient data transmission over the radio interface.
- **PDCP**: The eNodeB handles PDCP functions, which involve compression, encryption, and integrity protection of user data. PDCP ensures the secure and efficient transmission of user data packets over the air interface.
- **Radio Bearers Control**: The eNodeB establishes, configures, and releases radio bearers, which are logical channels that provide communication paths between the UEs and the core network. It manages the QoS parameters associated with each radio bearer, ensuring that the required performance levels are maintained.
- **Connection Control**: The eNodeB establishes and releases connections between UEs and the core network. It manages the signaling procedures for call setup, call release, and mobility-related events.
- **Measurement and Reporting**: The eNodeB collects and processes measurement reports from UEs to monitor the quality of the radio link. It uses these reports for handover decision-making, interference management, and overall network optimization.
- **Synchronization**: The eNodeB maintains synchronization with neighboring cells to ensure seamless handovers and accurate timing for data transmission.
- **QoS Management**: The eNodeB enforces QoS policies and manages the allocation of network resources to ensure the required performance levels for different types of services. It prioritizes and controls the traffic to deliver a consistent and reliable user experience.

By migrating functions to the eNodeB, the control plane processing is moved closer to the radio access network, reducing the latency and signaling overhead between the RAN and the core network. This results in improved system performance and responsiveness. The consolidation of these functions in the eNodeB simplifies network deployment and operation while enhancing system performance and scalability.

8.3.2 Evolved Packet Core/EPC

The EPC is the core network architecture of LTE-A, replacing the traditional core network elements of GSM/GPRS. It introduces significant evolutions in functionality, flexibility, and scalability to meet the requirements of high speed mobile broadband communications.

8.3.2.1 MME

The Mobility Management Entity (MME) is a critical component of the EPC responsible for managing mobility-related functions. It handles key functions such as subscriber management, authentication, security, and mobility management. It acts as the control plane anchor for the UEs and coordinates the establishment, reconfiguration, and release of connections. The MME maintains the location information of the UEs and tracks their movements, ensuring seamless handovers and efficient session management. It interacts with the eNodeBs over the S1-MME interface, exchanging control plane messages to establish and maintain the connection between the UEs and the core network. Additionally, the MME interacts with the HSS for subscriber authentication and authorization.

8.3.2.2 HSS

The HSS is a central database in the EPC that stores subscriber-related information. It holds the subscription profile, including the subscriber's authentication credentials, service profiles, and mobility management parameters. The HSS acts as the authentication center for the MME, providing the necessary authentication and authorization information to establish secure connections with UEs. It also plays a crucial role in subscriber management, handling registration and authentication processes. The HSS interacts with the MME over the S6a interface, providing authentication and authorization information for the UEs.

8.3.2.3 S-GW

The Serving Gateway (S-GW) is responsible for the routing and forwarding of user data packets in the EPC. It acts as the interface between the eNodeB and the external packet-switched networks. The S-GW performs packet routing and forwarding functions, ensuring efficient delivery of user data packets to the appropriate destination. It also plays a role in mobility management, handling handovers between eNodeBs and managing user sessions. The S-GW interacts with the eNodeB over the S1-U interface for the transport of user data packets. It also interfaces with the Packet Data Network Gateway (P-GW) over the S5/S8 interface for mobility-related procedures.

8.3.2.4 P-GW

The P-GW is responsible for providing connectivity between the LTE network and external packet-switched networks, such as the IP networks or the internet. It serves as the gateway for user traffic, performing functions such as IP address allocation, policy enforcement, and QoS management. The P-GW interfaces with the S-GW over the S5/S8 interface, exchanging control and user plane traffic. It also interfaces with the external packet-switched networks over the SGi interface, allowing for seamless connectivity and data transfer. The P-GW plays a crucial role in policy and charging control, enforcing policy rules and managing the charging of services.

8.3.2.5 PCRF

The Policy and Charging Resource Function (PCRF) is responsible for policy and charging control in the EPC. It is responsible for defining and enforcing policy rules that govern the behavior of network elements and UEs. The PCRF interacts with the P-GW over the Gx interface, providing policy and charging control decisions based on the subscriber's service profile and network conditions. It plays a crucial role in QoS management, ensuring that the network resources are allocated efficiently and in accordance with the defined policies. The PCRF also interfaces with the HSS for subscriber-specific policy control decisions.

Overall, the EPC provides a flexible and scalable core network architecture for LTE-A, enabling efficient mobility management, seamless connectivity, and enhanced service delivery. The migration of functions from the RNC to the eNodeB and the introduction of new network elements improve the performance, scalability, and capabilities of the LTE-A network.

8.4 LTE-A PHYsical Layer Enhancements

8.4.1 Enhanced Spectrum Efficiency

8.4.1.1 Baseband Radio

For a higher spectrum efficiency, LTE/LTE-A systems use much shorter symbols (66.68 µs) than UMTS does (about 665.6 µs). The radio frame structure is therefore also significantly different, as illustrated in Fig. 8.3. Every 10 ms frame consists of 20 0.5 ms slots, organized into 10 1 ms subframes. Depending on the specification of cyclic prefix, each slot carries either seven 66.68 µs symbols (with 5.21 µs short cyclic prefix), or six of them (with 16.67 µs extended cyclic prefix).

Though sharing the same overall frame structure, the uplink and downlink radio in LTE/LTE-A have significant differences in the control/data multiplexing, and in the allocation of reference signals. More specifically, in the uplink, control information elements are always intended to be either closest for the reference symbols in time domain or filled in the top rows of a resource block, as illustrated in Fig. 8.4. In the downlink, in comparison, the control information are always occupying the first symbols of each subframe (2–4 symbols for 1.4 MHz operation bandwidth, or 1–3 symbols for bandwidths above). Reference signals are distributed over the radio grid as illustrated in Fig. 8.5 and detailed specified in 3GPP TS 36.211 (2013). The data processing chains of the uplink SC-FDMA signals and of the downlink OFDM signals, as shown in Figs. 8.6 and 8.7, respectively, are generally similar. The only significant difference is an extra stage of transform precoding in the uplink, which serves for reducing the PAPR and therewith improving the power efficiency.

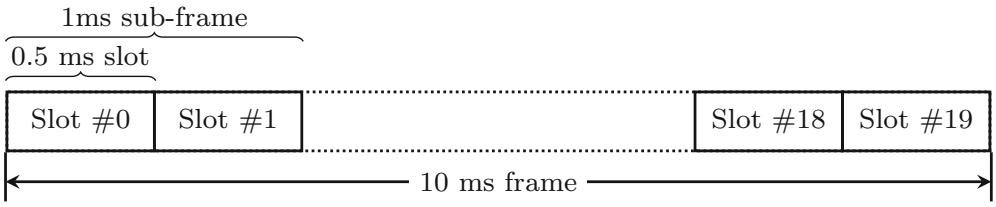


Fig. 8.3 The LTE/LTE-A FDD frame structure

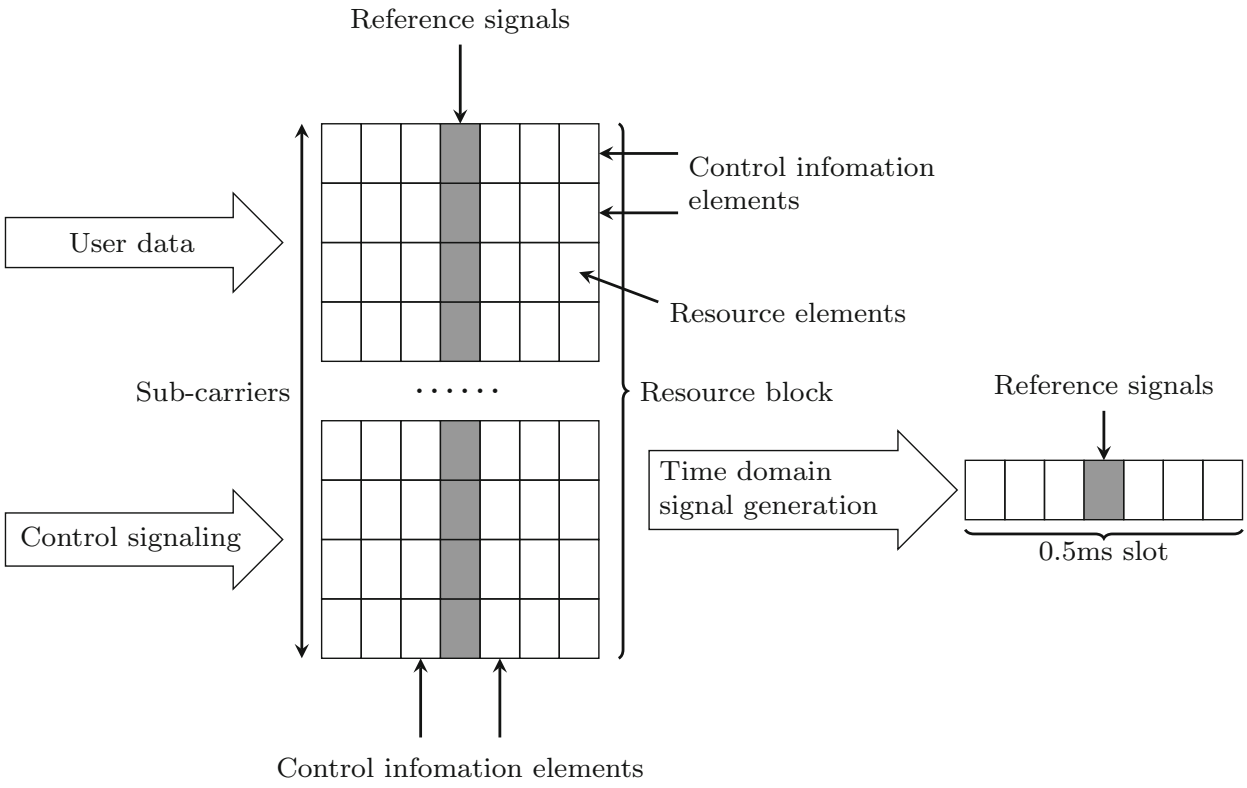


Fig. 8.4 The multiplexing of uplink control and data in LTE/LTE-A

8.4.1.2 Carrier Aggregation

In E-UTRA, CA is designed to operate in bands listed in Tables 8.2 and 8.3.

8.4.1.3 SU-MIMO and MU-MIMO

As we have introduced in Sects. 7.3.1 and 7.3.2, LTE-A introduces single-user multiple-input multiple-output (SU-MIMO) and multi-user multiple-input multiple-output (MU-MIMO) to support simultaneous transmission of multiple spatial layers between the E-UTRAN and one or multiple UEs, and therewith enhance the spectrum efficiency. More specifically, 2×2, 4×2, 4×4 and 8×2 MIMO are supported in the downlink, while 2×2, 2×4, and 4×4 are supported in the uplink. To fully exploit the potentials of the MIMO technologies, accurate CSI and flexible transmission mode is required. Therefore, LTE-A uses three different kinds of reference signals (RSs) in the downlink, namely the Cell Specific RS, the demodulation RS, and the channel state information reference signal (CSI-RS), to support CSI measurement and demodulation in different transmission modes.

While MU-MIMO can significantly enhance the data transmission rate, it generates also significant signaling overhead. Concerning this trade-off, in LTE-A Release 10, MU-MIMO is limited in dimension, more specifically: (i) no more than 4 UEs are co-scheduled; (ii) no more than 2 layers are allocated per UE; and (iii) no more than 4 layers are transmitted in total.

Moreover, to maximize the benefits of both SU-MIMO and MU-MIMO, LTE-A introduces dynamic switching between the two modes, allowing to change the mode per subframe, w.r.t. the channel condition and traffic.

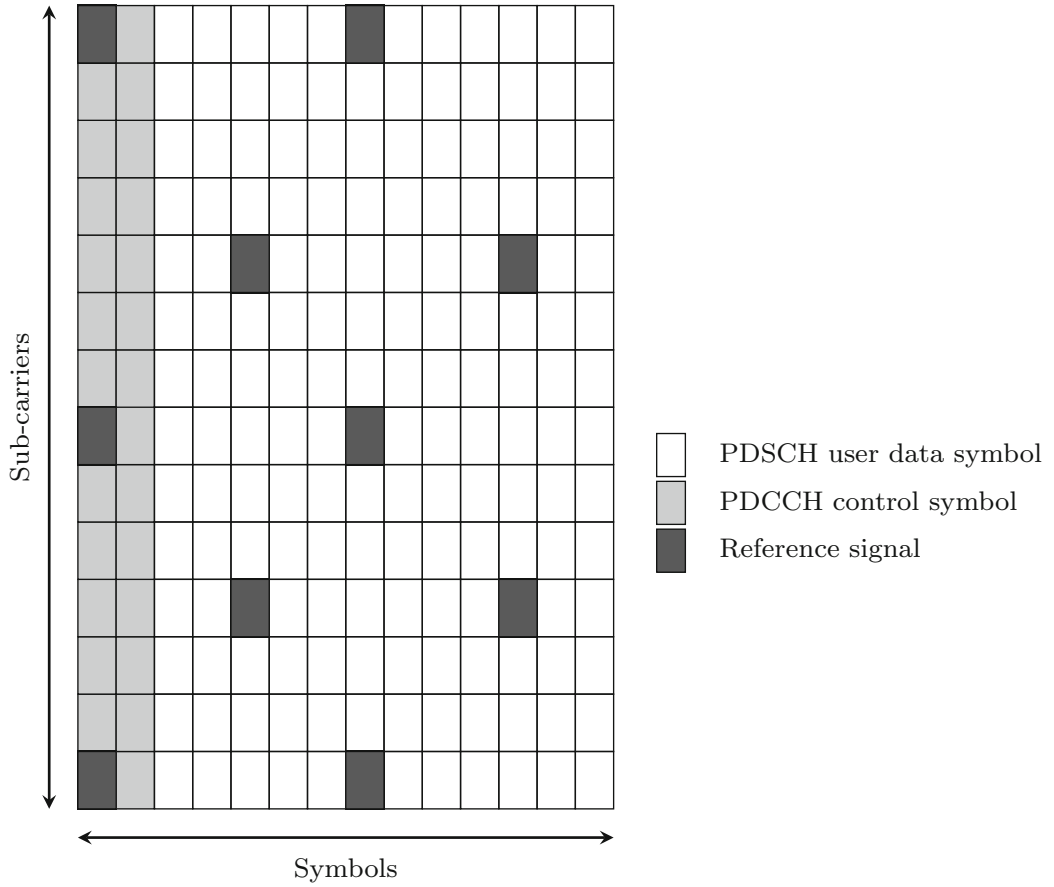


Fig. 8.5 The multiplexing of downlink control and data in LTE/LTE-A

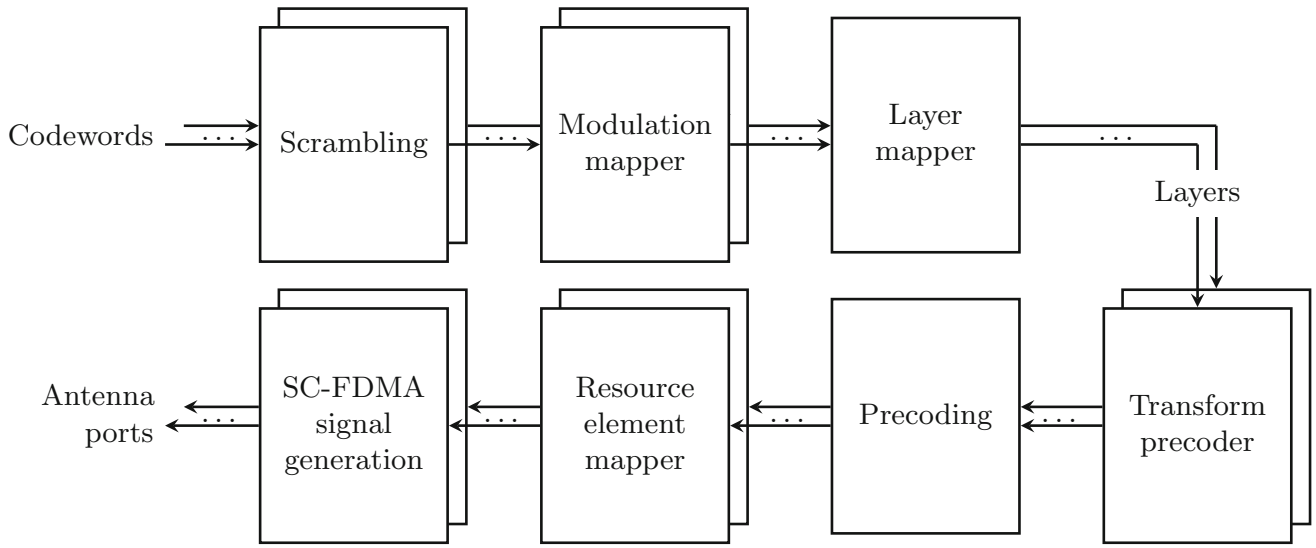


Fig. 8.6 The LTE/LTE-A PUSCH processing chain

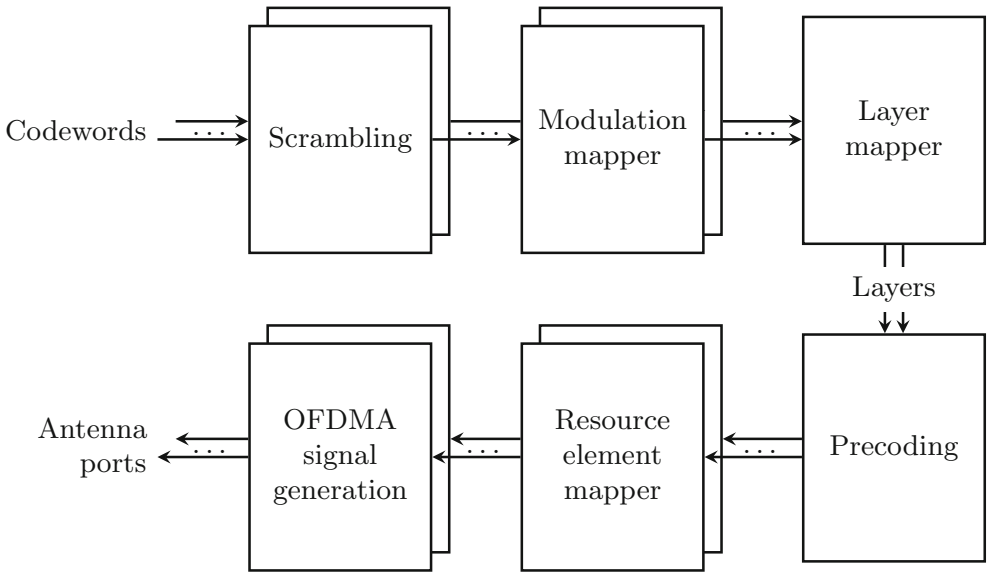


Fig. 8.7 The LTE/LTE-A downlink PHYsical channel processing chain

Table 8.2 Intra-band contiguous CA operating bands (3GPP TS 36.101 2022)

E-UTRA CA band	E-UTRA band	UL frequencies (MHz)	UL frequencies (MHz)	Duplex mode
CA_1	1	1920–1980	2110–2170	FDD
CA_40	40	2300–2400	2300–2400	TDD

Table 8.3 Inter-band contiguous CA operating bands (3GPP TS 36.101 2022)

E-UTRA CA band	E-UTRA band	UL frequencies (MHz)	UL frequencies (MHz)	Duplex mode
CA_1-5	1	1920–1980	2110–2170	FDD
	5	824–849	869–894	

8.4.2 Improved Interference Management

To mitigate inter-cell interference, which is a common challenge in cellular networks due to the frequency reuse among neighboring cells, LTE-A introduced ICIC.[1] By coordinating the resource allocation and transmission parameters, ICIC aims to improve the overall system performance and user experience. The ICIC functionality is located in the eNodeB.

Two ICIC modes were introduced in Release 10, namely fractional frequency reuse (FFR) and soft frequency reuse (SFR), using different ways to allocate and share frequency resources among neighboring cells. With FFR, the frequency spectrum is divided into different frequency subbands and allocated to different cells in a coordinated manner. The subbands are shared among neighboring cells, and each cell uses a subset of the available subbands for communication. Typically, the central region of each cell, known as the inner zone, is allocated a set of subbands that are not shared with neighboring cells, providing high quality and interference-free communication. The outer regions of the cell, known as the outer zones, share a different set of subbands with neighboring cells, allowing for better interference management. Therewith, FFR allows for more flexible and efficient allocation of frequency resources, especially in scenarios with varying cell sizes and different interference conditions, and achieves a balance between mitigating interference and achieving good frequency utilization. In comparison, SFR assigns different transmit power levels to different regions, in order to mitigate interference between neighboring cells, particularly at the cell edges, where interference tends to be more significant. Typically, the inner zone is allocated a higher transmit power, allowing for better signal quality and higher data rates, while the outer zones are allocated lower transmit power to reduce interference with neighboring cells. Therewith, SFR helps to improve system capacity and overall spectral efficiency by effectively managing inter-cell interference.

[1] The extended version of it, eICIC, was also introduced in LTE-A Release 10 as a general concept to address inter-cell interference challenges. However, it was not the primary focus of Release 10. The practical specifications and detailed implementation guidelines for eICIC were further developed and specified in subsequent releases, specifically in Release 12.

8.4.3 High Data Rates

To meet the high data rate requirements proposed in IMT-Advanced, LTE-A introduced enhanced MCS to support high modulation rates up to 64-QAM in both PDSCH and PUSCH (while UMTS supports up to 16-QAM). Meanwhile, the channel bandwidth was also increased from 5 MHz in UMTS to 60 MHz, or up to 100 MHz with CA (see Sects. 8.1 and 8.4.1.2). In addition, to achieve a satisfactory reception performance under high data rate, enhancements were made to the receiver design in various aspects.

First, the new reference signal design of LTE-A provides better potential for accurate channel estimation compared to legacy systems such as UMTS. This enables deployment of sophisticated algorithms to estimate the channel characteristics, such as fading and multipath effects, more accurately. Commonly applied estimating algorithms are including the Least Mean Square (LMS) and Minimum Mean Square Error (MMSE) (Kavitha & Manikandan 2015). Such accurate channel estimates enables the OFDM-based LTE-A systems to carry out frequency-domain equalization, providing up to 70% gain in the downlink efficiency compared to HSPA Release 6 (Holma & Toskala 2011).

Moreover, LTE-A incorporates interference rejection techniques to mitigate the effects of co-channel interference from neighboring cells or overlapping transmissions. These techniques include advanced interference cancelation algorithms, spatial filtering, and interference rejection combining (IRC) schemes. By mitigating interference, the receiver can improve the SNR and enhance the overall system capacity and performance.

8.4.4 Coverage Enhancements

8.4.4.1 Relay Nodes

As previously introduced in Sect. 7.3.6, relaying was standardized in LTE-A Release 10 as a key enabling technology to resolve coverage problems. In the relaying scenario, a RN is connected as the intermediate node between the UE and the so-called Donor eNodeB (DenB), which takes care of the data connection toward the core network. The protocol stack of the relay architecture is illustrated in Fig. 8.8.

In LTE-A, a RN may operate in two modes, namely (i) *inband*, where the UE-RN link and the RN-DenB link share the same carrier frequency, and (ii) *outband*, where the two links operate at different carrier frequencies. Depending on the duplex mode and whether holding their own cell IDs, RNs can be further categorized into four subclasses, as listed in Table 8.4.

8.4.4.2 HetNets

Heterogeneous Network (HetNet) is another important 4G enabling technology that enhances the coverage, as we have discussed in Sect. 7.3.10. In LTE-A, a HetNet typically consists of MACro cells, which are large coverage cells deployed in

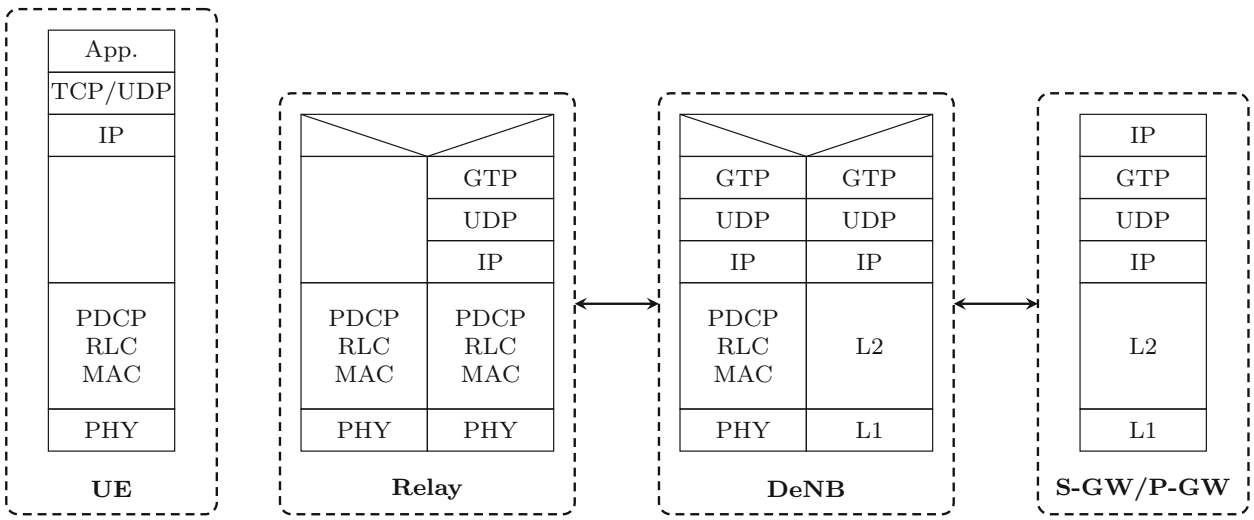


Fig. 8.8 Relay architecture

Table 8.4 Relay classes in LTE-A Release 10

Class	Cell ID	Mode
Type 1	yes	inband half-duplex
Type 1a	yes	outband full duplex
Type 1b	yes	inband full duplex
Type 2	no	inband full duplex

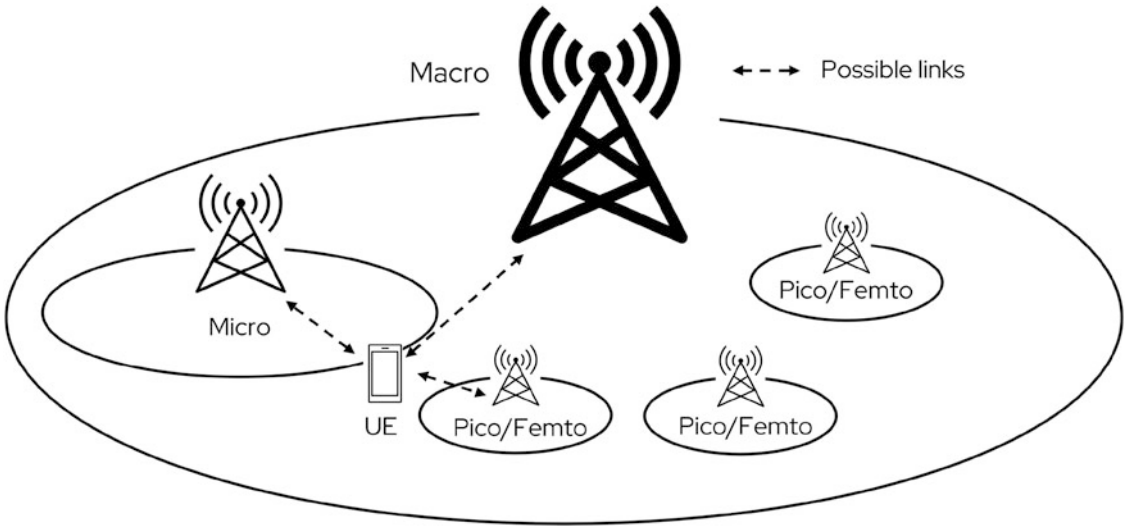


Fig. 8.9 Typical heterogeneous deployment of LTE-A network

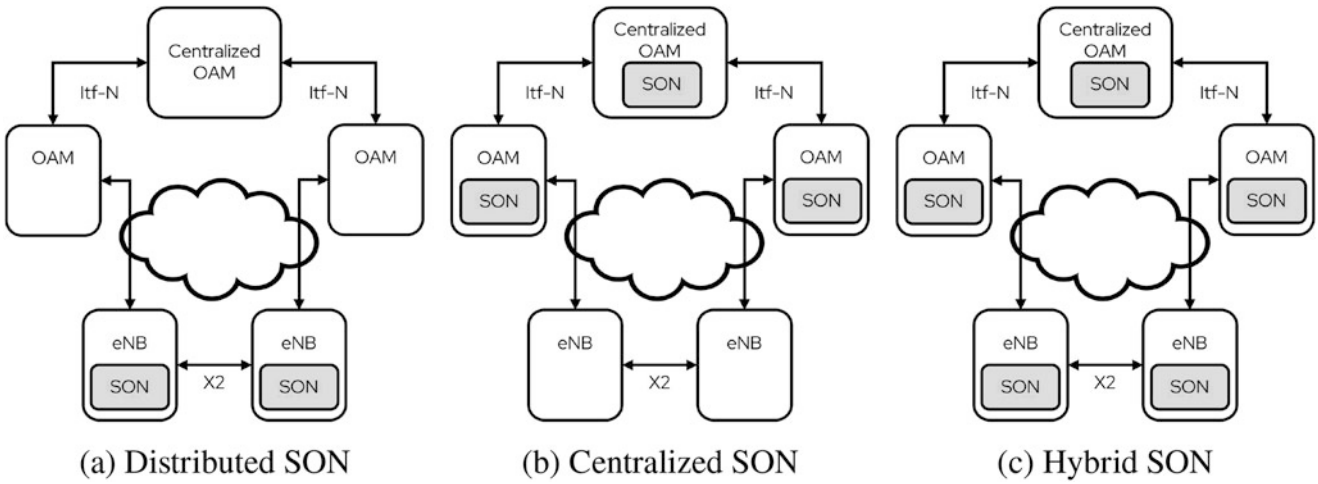


Fig. 8.10 SON architecture options. (**a**) Distributed SON. (**b**) Centralized SON. (**c**) Hybrid SON

a traditional manner, and small cells (including microcells, pico cells, and femto-cells), as illustrated in Fig. 8.9. Small cells are deployed in a denser fashion to enhance network capacity and coverage in specific areas such as indoor environments, busy urban areas, or hotspot locations.

8.4.4.3 SON

In addition, SON, as previously introduced in Sect. 7.3.13, further improves the coverage performance of LTE-A by means of enabling self-configuration, self-optimization, and self-healing of the network. The three deployment options, i.e., distributed, centralized, and hybrid (a.k.a. localized) SON, are illustrated in Fig. 8.10.

8.5 LTE-A Radio Interface Protocols

8.5.1 Simplified Protocol Stack

Compared to UMTS, LTE/LTE-A significantly enhanced its radio interface protocol stack, as depicted in Fig. 8.11, to streamline operations and increase efficiency. A key modification is the removal of the BMC sublayer from L2, which simplified the architecture by eliminating redundancies. The functionalities of BMC, such as managing paging messages, were effectively absorbed by other layers like RRC and NAS, thereby reducing complexity and enhancing protocol processing times. Additionally, the PDCP sublayer was expanded to include control plane functions. This expansion enabled PDCP to manage security functions - previously handled by RLC in UMTS - and header compression for both user data and control plane data, thus enhancing data transmission security and efficiency. These changes not only simplified the protocol stack but also improved resource management and data processing speeds, aligning with LTE/LTE-A's objectives of achieving higher data rates and lower latency.

A more specific illustration of the radio interface protocols in the LTE/LTE-A architecture is given by Fig. 8.12.

8.5.2 Radio Resource Management

In LTE/LTE-A, various RRM functions are on different protocol layers, from PHY to PDCP and RRC. Their implementation in eNodeB is listed in Table 8.5.

With PHY features discussed in Sect. 8.4, here below we mainly focus on Layers 2 and 3.

8.5.2.1 HARQ

Similarly to HSDPA and HSUPA, LTE/LTE-A uses a pipeline of multiple stop-and-wait HARQ processes on the MAC sublayer to maintain a continuous data stream with minimized overhead. More specifically, the number of HARQ processes

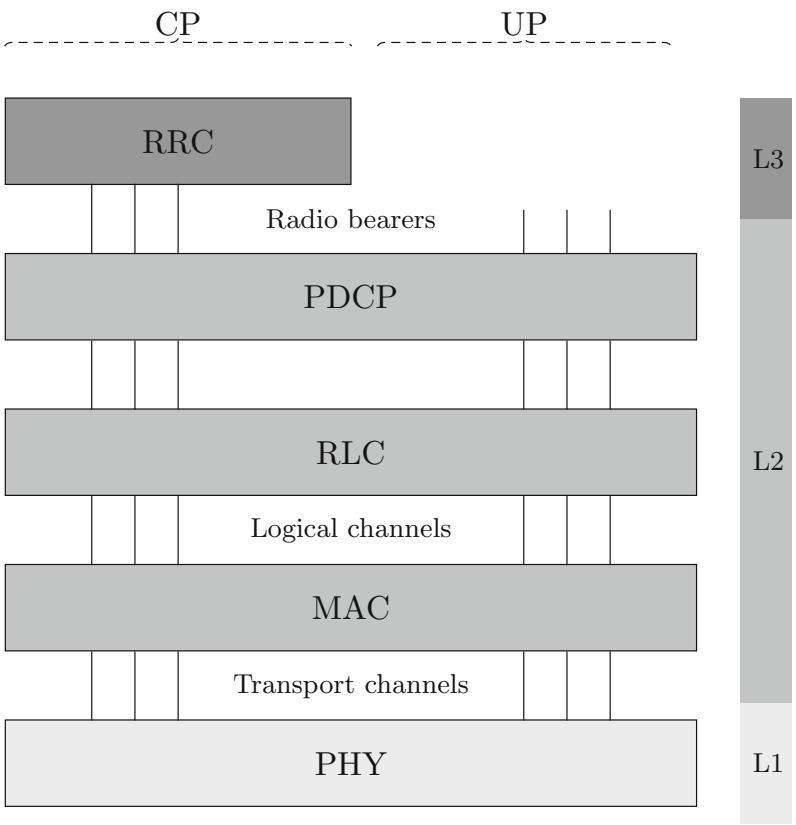


Fig. 8.11 LTE/LTE-A radio interface protocol stack

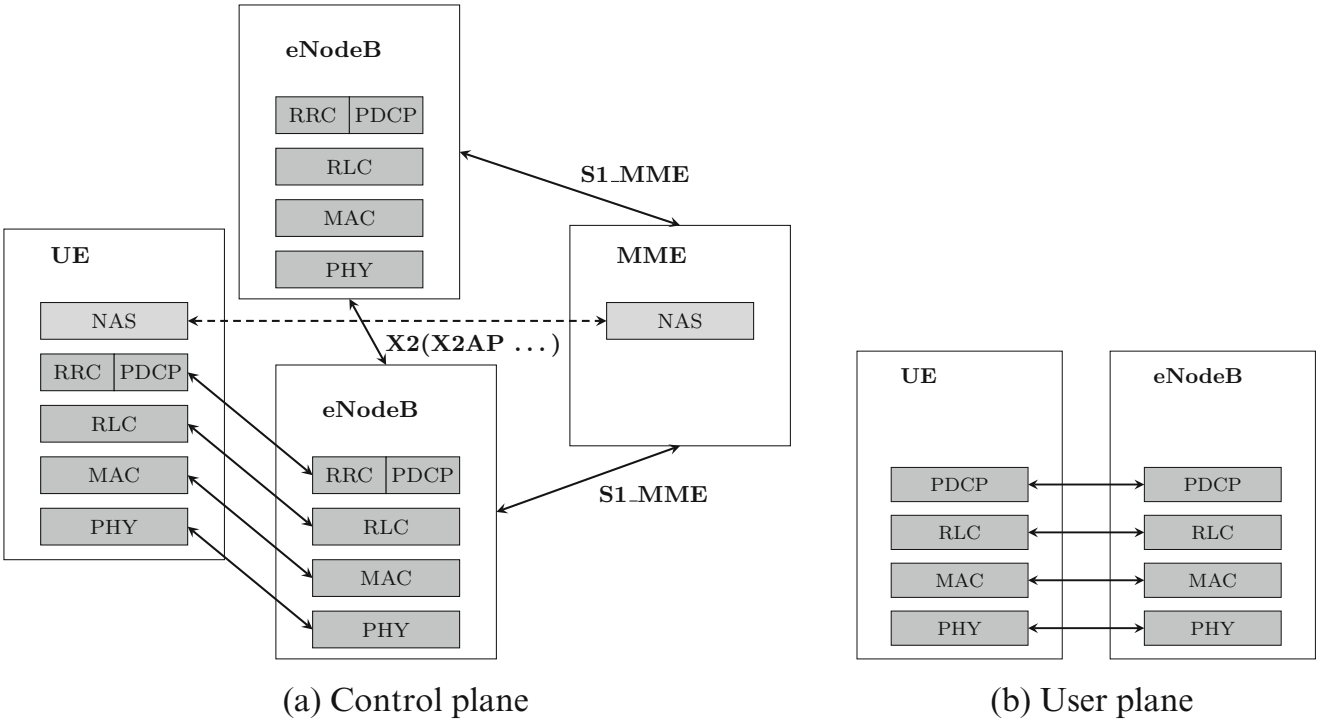


Fig. 8.12 LTE/LTE-A radio interface protocols. (**a**) Control plane. (**b**) User plane

Table 8.5 RRM functions in eNodeB

	User plane	Control plane	RRM functions
L3	PDCP[a]	RRC	QoS management
			Admission control
			Persistent scheduling
L2	RLC	RLC	HARQ manager
	MAC	MAC	Dynamic scheduling
			Link adaptation
L1	PHY	PHY	PDCCH adaptation
			CQI manager
			Power control

[a] As commented in Sect. 6.4, the 3GPP protocol stack is not perfectly aligned with the 7-layer OSI model. While PDCP is considered as part of Layer 2 in most contexts, it can also be classified as part of Layer 3, particularly when referring to RRM

is fixed to 8 in LTE/LTE-A for both UL and DL (in UMTS the number is 4). Taking the downlink as example, each Physical Downlink Shared Channel (PDSCH) packet lasts 1 ms, so as the ACK/NACK packet over PUCCH or PUSCH. With 8 HARQ processes, the interval between two packets/retransmissions in each process is 8 ms, leaving 3 ms processing time for each side of the link, i.e., the UE and the eNodeB. Both chase combining and incremental redundancy are applicable for the soft combining.

In addition, if a packet fails to be decoded despite of the HARQ mechanism on the MAC sublayer, an ARQ procedure will be triggered on the RLC sublayer.

8.5.2.2 Dynamic Scheduling and Link Adaptation

Downlink

Unlike the CDMA-based UTRAN, the downlink packet scheduling in OFDMA-based E-UTRAN is carried out not only in the time domain but also in the frequency domain. However, a full time-frequency joint scheduling can be computationally complex, leading to high latency and/or cost. Therefore, LTE-A applies a three-step packet scheduling in the downlink: a time-domain scheduling is first carried out, which selects up to N users for potential scheduling during the next TTI. Then, it

executes a control channel schedule check to verify if enough resources are available in the PDCCH for all selected users—if not, upon the priority, a subset of the users will be excluded. The remaining selected users are then scheduled by the frequency-domain scheduler for physical resource block (PRB) allocation. Especially, the frequency-domain scheduler is HARQ aware, i.e., it prioritizes the users with first transmissions of new data with the best PRBs, and allocates the remaining PRBs for users with pending HARQ retransmissions.[2] Furthermore, to optimize the packet scheduling, channel-specific information such as supported MCS of a user upon the selected PRBs is provided by the link adaptation unit, which consists of an outer loop link adaptation (which is similar to the algorithm used in HSDPA), an inner loop adaptation, and a MIMO adaptation algorithm. The link adaptation decisions are primarily based on the CQI feedback from users.

Uplink

The packet scheduling in uplink exhibits several key differences when compared to that in downlink. First of all, since the eNodeB does not possess full knowledge about the buffered data amount at UE by nature, additional signaling is required to report this buffer status and support the decision. Second, the power budget is generally less in UL than in DL due to the limited battery capacity of UE, which constraints the PRB allocation for users with poor channels. Third, only adjacent PRBs can be allocated to the same UE in the uplink due to the SC-FDMA scheme. Furthermore, the interference is typically more dynamic in the UL than in the DL.

Similar to the downlink case, in the uplink the link adaptation unit also collaborates closely with the packet scheduling. The UL link adaptation consists of an open-loop link adaptation, an adaptive modulation and coding, as well as a power control algorithm. Its decisions are based on the SINR measurements and the CSI reports.

Moreover, like UMTS, LTE/LTE-A allows UEs to save power by means of working in the discontinuous reception (DRX)/DTX mode. However, once a UE goes into this mode, it stops from listening to the L1/L2 control channel and will be no more scheduled for transmission on PUSCH. Regarding this, the packet scheduling also receives DRX status information from the DRX management unit as an essential input.

8.5.2.3 QoS Management and Admission Control

On Layer 3, LTE/LTE-A provides enhanced QoS management and admission control, which closely collaborate to provide support for various services and applications with diverse QoS requirements.

Each Evolved Packet System (EPS) bearer is associated with a QoS profile with several parameters, depending on the type of bearer, the involved QoS parameters can be different, as listed in Table 8.6.

Generally, bears can be classified into default and dedicated ones. A default bearer will be assigned to every UE when it attaches to the LTE/LTE-A network for the first time, and will remain as long as the UE is attached. Each default bearer comes with a specific Access Point Name (APN) and an IP address. When multiple default bearers are assigned to the same UE, which is supported by LTE, each of them will have a separate IP address. Every default bearer is also specified with an allocation retention priority (ARP), for deciding whether new bearer modification or establishment request should be accepted considering the current resource situation. Default bearers are best-effort service, so they are not specified with any guaranteed bit rate (GBR) or dedicated maximum bit rate (MBR), but with an APN-aggregate MBR (AMBR) for the maximum allowed throughput to the specific APN, and a UE-AMBR for the maximum allowed throughput among all APN to the specific UE. Both AMBRs are specified interdependently for the uplink and the downlink. Dedicated bearers, in comparison, are used to provide dedicated tunnels to specific traffic (e.g., conversational video, real-time gaming, etc.). Each dedicated bearer acts as an addition on top of a default bearer, with which it shares the APN and IP address. Each dedicated bearer is associated with a Linked EPS bearer ID (ID) and a traffic flow template (TFT). The former identifies the default bearer to which the dedicated bearer is attached, and the latter defines traffic rules that which IP packet should be sent on the particular dedicated bearer.

Dedicated bearers can be further divided into GBR or non-GBR ones. GBR bearers are supposed to deliver guaranteed bit rate for QoS-critical services, so they are specified with GBR and MBR. Non-GBR dedicated bearers, on the other hand, are similar to default bearers that they provide only best-effort services, and are therefore only specified with AMBRs.

Every EPS bearer is assigned with a QCI as part of its QoS profile. There are 9 different QCIs defined, where QCIs 1–4 are applicable for GBR bearers, and 5–9 for non-GBR ones. Detailed specifications are listed in Table 8.7.

Compared to the QoS concept in UMTS, which contains more than 10 different QoS parameters, LTE/LTE-A has a simplified QoS profile for better signaling efficiency. Meanwhile, it provides more QCI to support more classes of services with diverse QoS requirements.

[2] In LTE/LTE-A, it is impossible to simultaneously schedule new data and pending HARQ retransmissions for the same user in the same TTI.

Table 8.6 LTE-A QoS parameters

Bearer type		QoS parameters
Default		QoS class identifier (QCI) 5–9
		APN
		APN—AMBR
		UE—AMBR
		IP address
		ARP
Dedicated	Non-GBR	QCI 5–9
		APN—AMBR
		UE—AMBR
		TFT
		ARP
		L-EBI
	GBR	QCI 1–4
		GBR
		MBR
		TFT
		ARP
		L-EBI

Table 8.7 QCI characteristics for the EPS bearer QoS profile

QCI	Bearer type	Priority	Packet delay budget (L2)	Packet loss rate (L2)	Example services
1	GBR	2	100 ms	10^{-2}	Conversational voice
2	GBR	4	150 ms	10^{-3}	Live streaming
3	GBR	3	50 ms	10^{-3}	Real-time gaming
4	GBR	5	300 ms	10^{-6}	Buffered streaming
5	non-GBR	1	100 ms	10^{-6}	IMS signaling
6[a]	non-GBR	6	300 ms	10^{-6}	Buffered streaming, email, file sharing, etc.
7	non-GBR	7	100 ms	10^{-3}	Voice, live streaming, interactive gaming
8[a]	non-GBR	8	300 ms	10^{-6}	Buffered streaming, email, file sharing, etc.
9[a]	non-GBR	9	300 ms	10^{-6}	Buffered streaming, email, file sharing, etc.

[a] QCI 6 could be used for the prioritization of non-real-time data if the network supports Multimedia Priority Services (MPS). QCIs 8 and 9 are typically used for dedicated bearers for "premium subscribers," and default bearers for "nonprivileged subscribers," respectively

The admission control algorithm in eNodeB makes decision whether requests for new EPS bearers in the cell shall be granted or rejected, according to their QoS profiles and the resource situation. New requests can only be admitted when the associated QoS requirements can be fulfilled.

8.5.2.4 Persistent Scheduling

In LTE-A, the persistent scheduling on Layer 3 differs from the L2 dynamic scheduling. While the latter is making real-time decisions of PRB allocation in time and frequency domains, as we have introduced in Sect. 8.5.2.2, the former is a static RRM approach designed to support low-latency, real-time communication for certain types of services. It allocates a dedicated set of resources to specific UEs continuously and persistently over multiple scheduling intervals, even when they have low or no traffic during certain intervals. The persistent allocation ensures that the resources are instantly available whenever the UEs need to transmit data, eliminating the delay associated with dynamic resource assignment. While dynamic scheduling is more flexible and adaptive, allowing resources to be allocated to UEs based on their varying needs and network conditions, persistent scheduling, on the other hand, guarantees low-latency access to resources for specific UEs, making it suitable for delay-sensitive services like real-time voice or video calls.

8.6 LTE/LTE-A Mobility Management

The mobility management LTE/LTE-A represents a significant improvement over the UMTS, primarily due to simplification in RRC states and enhanced procedures for cell selection/reselection, handovers, and tracking area optimization.

8.6.1 Simplified RRC States

Compared to UMTS which has five different RRC states (see Fig. 6.29), LTE/LTE-A simplified this by having only two states: RRC_IDLE and RRC_CONNECTED. This simplification minimizes the number of state transitions, resulting in lower latency, improved responsiveness, and efficient use of resources. In RRC_IDLE, the UE performs cell selection/reselection, while in RRC_CONNECTED, the UE engages in data transfer.

8.6.2 Mobility Management in Idle State

In LTE/LTE-A, UEs in idle state are limited with little functionality but location registration, PLMN selection, and cell (re)selection, as shown in Fig. 8.13.

Cell selection and reselection processes are of paramount importance in idle state mobility management. It is autonomously performed by the UE to ensure the highest quality of connection by choosing the best cell based on UE measurements and network-broadcast parameters. Basically, a cell is only considered suitable by the UE, when its reference signal receive power (RSRP) measured by the UE exceeds a predefined minimum required received level, which can be optionally adjusted by an offset when searching for a higher priority PLMN.

When multiple cells fulfill the criteria, cell (re)selection can be performed based on cell ranking. The most basic case is the intra-frequency and equal priority reselection, in which the cells are equally ranked and compared based on their RSRPs. To avoid ping-pong phenomena, an extra reselection RSRP margin Q_{hyst} and a timeout $T_{reselect}$ are applied, as illustrated in Fig. 8.14. When it comes to inter-frequency and inter-RAT reselections, LTE/LTE-A provides an advanced procedure to optimize the selection process. Each specific RAT/frequency is called in this procedure a *layer*, and each layer is given a certain priority. Every UE always tend to camp on the highest priority layer that its RSRP can fulfill the requirement. For reselection, two thresholds are specified for every individual layer, to support the decision if the UE is reselecting toward (respectively from) a higher priority layer than the current serving one.

It is worth noting that the new X2 interface introduced in LTE enables direct communication between eNodeBs, making the reselection process more efficient.

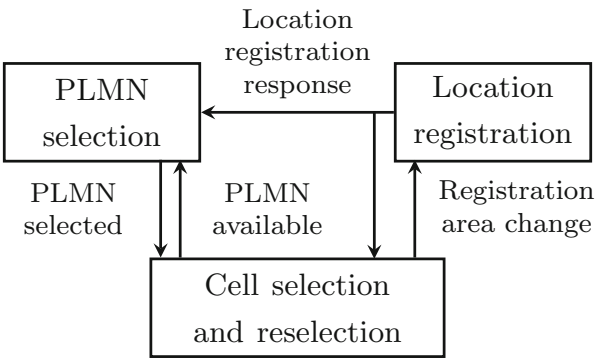


Fig. 8.13 Overview of LTE/LTE-A idle mode (Holma & Toskala 2011)

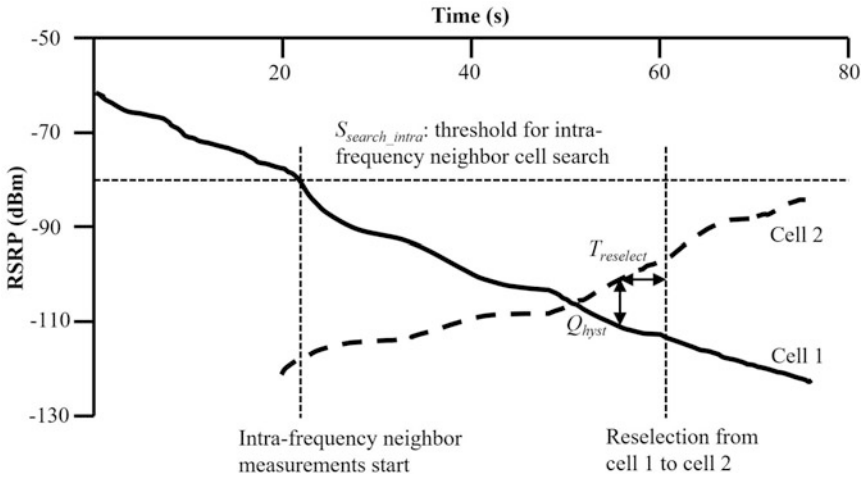


Fig. 8.14 Intra-frequency cell reselection in idle mode (Holma & Toskala 2011)

8.6.3 Tracking Area Optimization

The concept of a tracking area (TA) in LTE/LTE-A is an evolution of the routing area (RA) in UMTS, designed to more efficiently manage the mobility of UEs. A UE typically belongs to one TA at a time, and it notifies the network when it moves to a different TA through a procedure known as a TA Update TAU, which is similar to the LU and RA Update in legacy systems.

One major enhancement in TA and TAU of LTE/LTE-A is that it allows multiple TAs be grouped into a list, known as the TA List (TA). Each UE can be assigned with an individual TAL based on its mobility pattern, and the TAU procedure is not triggered when the UE moves around different TAs on its current TAL. This approach optimizes the TAU procedure by reducing the frequency of updates, especially for UEs with high-mobility patterns.

8.6.4 Intra-LTE Handover

The intra-LTE handover procedure is illustrated in Fig. 8.15. Unlike the RNC-controlled soft handovers in UMTS, LTE/LTE-A employs hard handovers, which are directly managed by the eNodeB. The eNodeB leverages the X2 interface to coordinate handovers directly with other eNodeBs, resulting in reduced handover latency and better user experience. During the radio handover, the user plane traffic flows in the packet forwarding mode, like shown in Fig. 8.16: the UL is directly switched to the target eNodeB, while the DL is first forwarded by the source eNodeB, over the target eNodeB through the X2 interface, until the completion of handover. The detailed signaling message flow is illustrated in Fig. 8.17.

8.6.5 Inter-RAT Handover

In the same way like UTRAN provides backward-compatibility with GERAN, E-UTRAN also supports inter-RAT handovers from/to UTRAN and GERAN. Inter-RAT handovers are designed in backward manner that the decisions are made by the source system, and the radio resources are reserved in the target system before the handover command is issued to the UE. Exceptionally, since GERAN does not support PS handover, it does not reserve the resources before the handover. Since there is no direct interface between RANs of different RATs, the signaling is performed over the core network, as illustrated in Fig. 8.18.

With the inter-RAT handover, together with inter-RAT reselection, the UE may experience transitions between different RRC states across different RATs, as briefly summarized in Fig. 8.19.

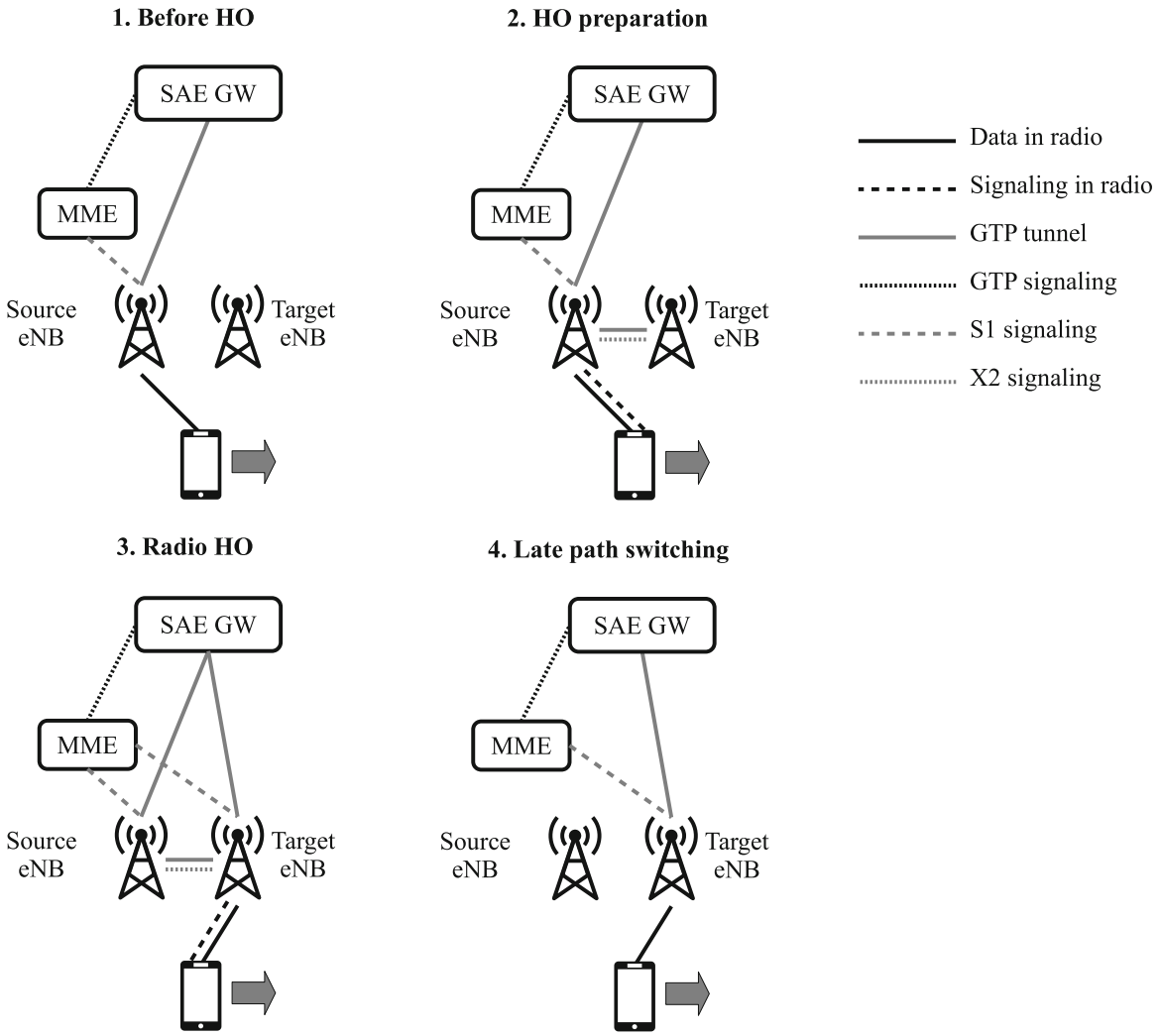


Fig. 8.15 Intra-LTE handover procedure

8.6.6 Summary of Differences in Mobility Management between LTE-A and UMTS

In conclusion, LTE/LTE-A has brought significant advancements regarding mobility management compared to UMTS. Simplification of RRC states, enhancements in idle state MM, optimization of tracking areas, and improvements in handover procedures all contribute to improved system performance, network efficiency, and overall user experience. A summary of the main differences between the mobility management in E-UTRAN and that in UTRAN is outlined in Table 8.8.

8.7 LTE/LTE-A Security

While inheriting most of the security architecture from UMTS, LTE/LTE-A has also brought significant advancements in security, as summarized in Table 8.9.

8.7.1 Unique Threats to LTE Networks

Despite the advancements in security, LTE/LTE-A networks are not immune to threats. Some of the unique threats to LTE networks are including:

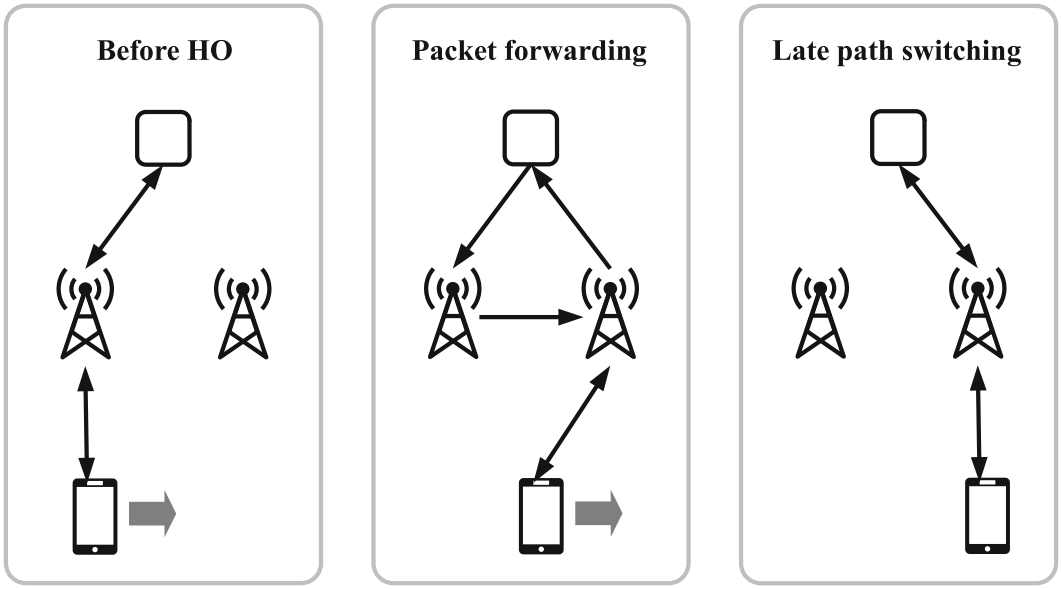


Fig. 8.16 User plane switching in intra-LTE handover

- **Renegotiation attacks**, which exploit the fact that LTE allows the renegotiation of security parameters during a session. An attacker can force a downgrade of the security parameters, making the network vulnerable to further attacks.
- **Jamming** of the UE radio interface, by which an attacker can prevent the UE from communicating with the network.
- **Availability attacks** on the eNodeB and core network are also a concern, which aim to disrupt the availability of network services, causing significant disruption to users.

8.8 Summary to LTE TDD Mode

Unlike UMTS where the FDD and TDD modes are separately specified, LTE/LTE-A standardizes both modes in a harmonized way that they are included together in a same set of specifications. The differences between the two duplex modes in LTE/LTE-A are few and mild, dominantly focusing on the PHY layer regarding the UL/DL switching, in addition to even fewer lying on the MAC and higher layers which are involved with the PHY layer parameters. The architecture and procedures of LTE/LTE-A are identical in FDD and TDD, making it possible to support a coexistence of the two modes within the same network, or even the same UE.

Compared to LTE-A FDD, LTE-A TDD shows the following advantages:

1. **Compatibility with TD-SCDMA**: the extensive deployment of TDD 3G systems in some countries, primarily the TD-SCDMA systems in China, urges to reuse the unpaired spectrum also in LTE/LTE-A in coexistence with the legacy systems. LTE-A TDD can be easily configured to achieve timing alignment on both the frame level and the subframe level (Holma & Toskala 2011), accommodating it to coexist with TD-SCDMA without interfering each other.
2. **Asymmetric UL/DL capacity allocation**: by adjusting the duplex switching point, LTE-A TDD is able to flexibly allocate the overall channel capacity to UL and DL frames, which can be dynamically updated.
3. **Channel reciprocity**: sharing the same frequency band, the UL and DL channels in TDD systems are highly correlated to each other, which allows to partially or fully use the channel measurement in one direction for the other. A reduction in signaling overhead can be therewith achieved.

However, LTE-A TDD also exhibits its drawbacks, mainly including:

1. **Inter-operator interference**: closely deployed LTE-A TDD systems by different operators will easily interfere each other, if they work in the same or adjacent bands without synchronizing their base stations with the same UL/DL switching timing. Therefore, coordinated deployment is recommended for LTE-A TDD, otherwise guard bands and additional filtering will be needed. Either way is leading to increased system complexity and therefore more cost.

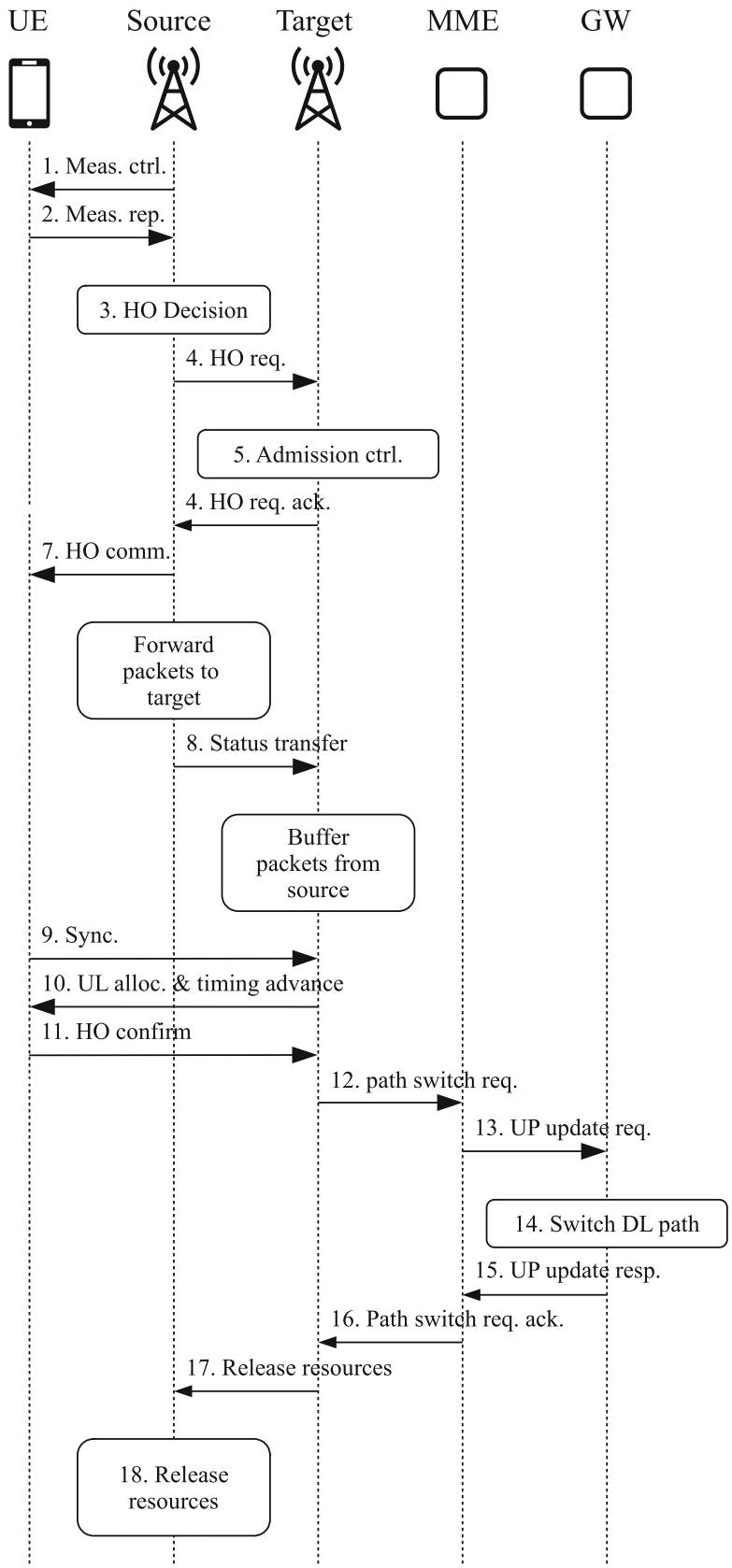


Fig. 8.17 Signaling of the intra-LTE handover procedure

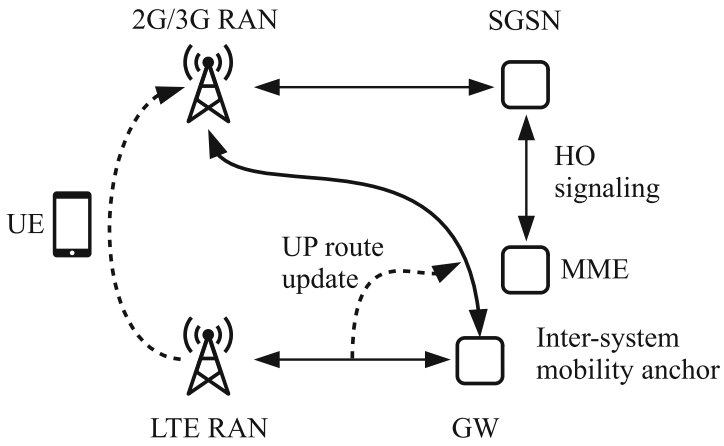


Fig. 8.18 Inter-RAT handover procedure from E-UTRAN to UTRAN/GERAN (Holma & Toskala 2011)

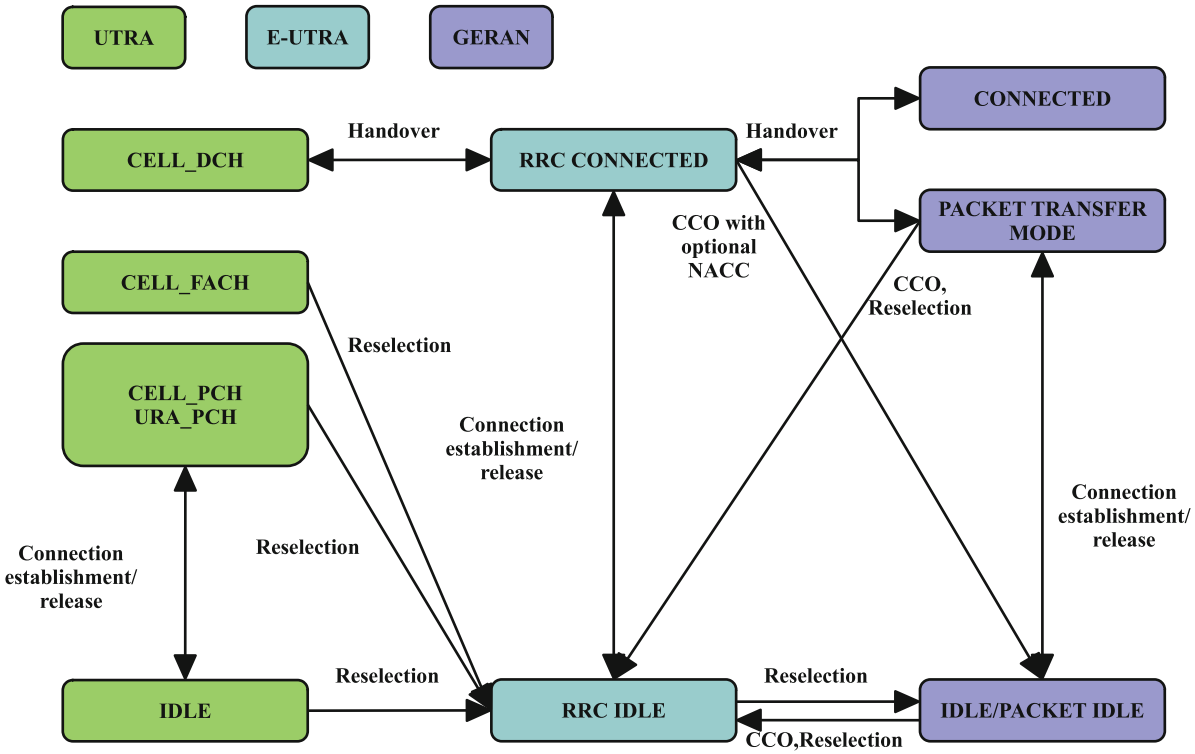


Fig. 8.19 RRC state transitions across GERAN, UTRAN, and E-UTRAN

2. **Limited cell size**: when the cell size increases, the maximal delay spread also grows, so that the guard period between UL and DL transmissions in LTE-A TDD must be correspondingly increased. This leads to a loss in the time efficiency and reduces the achievable data rate, especially in rural deployments where macro cells are commonly large.

8.9 IoT over LTE-A

Predominantly focusing on the traditional demand for human-type communication (HTC), LTE-A is also expected to serve in MTC scenarios such as IoT, which generally have lower requirements for peak data rate and throughput in comparison to HTC, but can bare significantly less power consumption. MTC over LTE has been possible since 3GPP Release 8, where 5 distinct UE categories are defined, each specified to fulfill the requirements of a typical use case. This list was

Table 8.8 Comparison between LTE-A and UMTS Mobility Management

UTRAN	E-UTRAN	Notes
LA (CS core)	Not relevant (no CS connections)	For CS fallback handover, MME maps TA to LA
RA	TA	
Soft handover used for WCDMA UL/DL and for HSUPA UL	No soft handover	
Cell_FACH, Cell_PCH	No similar RRC states	E-UTRAN always uses handovers for RRC connected users
RNC hides most of mobility	Core network sees handover	Flat architecture HSPA is similar to E-UTRAN
Neighbor cell lists required	No need to provide cell specific information, i.e., only carrier frequency is required, but the network can provide cell specific reselection parameters if desired	Also UTRAN UE can use detected cell reporting to identify a cell outside the neighborlist

Table 8.9 Comparison between LTE-A and UMTS Security

	UMTS	LTE/LTE-A
Authentication	Mutual authentication using AKA protocol.	Mutual authentication using EPS AKA protocol. Key Derivation Functions use HMAC-SHA-256
Integrity protection	Kasumi algorithm used, mandatory for only few RRC messages.	SNOW 3G and ZUC algorithms used, mandatory for all messages after (and including) Security Mode Command
Encryption	No encryption or Kasumi algorithm used.	Encryption done independently at two levels; two SECURITY MODE COMMANDS for two sets of keys; no encryption, SNOW 3G, or AES-128 algorithm used.
Backhaul protection	No specific requirement for backhaul protection.	Mandates the use of IPsec for backhaul protection.

Table 8.10 3GPP LTE solutions for IoT

	Cat 1	Cat 1bis	Cat 0	Cat M1	Cat M2	Cat NB1	Cat NB2
3GPP release	8	13	12	13	14	13	14
DL Peak Data Rate	10 Mbps	10 Mbps	1 Mbps	1 Mbps	∼4 Mbps	24 kbps	127 kbps
UL Peak Data Rate	5 Mbps	5 Mbps	1 Mbps	1 Mbps	∼7 Mbps	66 kbps	159 kbps
UE Bandwidth	1.4-20MHz	1.4-20MHz	1.4-20MHz	1.4 MHz	5 MHz	180 kHz	180 kHz
UE Transmission Power	23 dBm	23 dBm	23 dBm	20/23 dBm	20/23 dBm	20/23 dBm	14/20/23 dBm
Latency	<100 ms	<100 ms			<15 ms		<10 s
Antennas	2	1	1	1	1	1	1

then continuously extended throughout the evolution of LTE, reaching to a total number of 24 different UE categories in Release 14.

In Release 8, it is the Cat 1 focusing on IoT applications, which supports only one single DL MIMO layer, with peak data rates of 10.3 Mbps and 5.2 Mbps in the DL and UL, respectively. In comparison, the corresponding numbers for Cat 5 are 4 layers, 299.6 Mbps, and 75.4 Mbps, respectively. At this reduced data rate, Cat 1 UEs benefits from lower power consumption and lower implementation cost.

However, the standard LTE specifications generally require UE to have multiple antennas, which Release 8 failed to explicitly exclude Cat 1 UEs from. This can be unnecessary for low-cost IoT devices and therefore annoying the vendors. Upon the strong demand for single-antenna IoT hardware, 3GPP introduced in Release 13 the Cat 1bis, which is fundamentally Cat 1 except for that the UE uses single antenna.

Meanwhile, due to the blooming prosperity of the IoT market and the therewith rising demand, 3GPP further introduced two more cellular IoT solutions in Releases 12–14: the LTE-MTC a.k.a. LTE-M, represented by the UE categories 0, M1, and M2; and the NB-IoT, represented by Cat NB1 and Cat NB2, respectively.

An overview to the 3GPP LTE solutions for IoT is provided in Table 8.10.

8.10 Summary

This chapter delves into the 4G successor of UMTS: the LTE-A, based on 3GPP Release 10. In this chapter we provided a comprehensive overview of LTE-A, focusing on its key features, architecture, radio interface protocols, mobility management, and security. We highlighted the significant advancements and differences from its predecessor, UMTS. Focusing on Release 10, which is the first standard that fulfills the 4G requirements proposed in IMT-Advanced, this chapter does not cover technologies introduced in later releases of LTE-A, such as CoMP, D2D communication, LAA, etc.

8.11 Exercises

1. Describe the main components of the LTE-A system architecture and their functionalities.
2. How does the architecture of E-UTRAN differ from that of UTRAN?
3. Outline the key changes and enhancements in the radio interface protocol stack when transitioning from UMTS to LTE-A.
4. Discuss the functionalities of the PDCP, RLC, MAC, and PHY layers in the LTE-A protocol stack and how they evolved from their UMTS counterparts.
5. Discuss the significance of OFDMA and SC-FDMA in the LTE-A PHYsical layer.
6. How does MIMO enhance the performance of LTE-A networks?
7. Describe the purpose and benefits of Carrier Aggregation in LTE-A.
8. Explain the role of Relaying and Heterogeneous Networks in enhancing network coverage and capacity.
9. Describe the X2 interface's role in LTE-A and its significance in mobility management.
10. Explain the concept of "Inter-eNB handover" in LTE-A and how it differs from traditional handover mechanisms.
11. Discuss the EPS AKA mechanism in LTE-A and its advantages.
12. Explain the significance of NAS security in LTE-A and how it ensures secure communications.

9 Evolution to Fifth-Generation (5G) Mobile Cellular Communications

9.1 5G: From Connecting People to Connecting Things

The development of 5G technology was spurred by emerging user and application trends, the exponential growth in mobile traffic, and significant technological advancements that made high-frequency spectrum feasible. Research on 5G began early, even before the full deployment of 4G, taking inspiration from the success of LTE technology. In the early 2010s, some efforts were made to explore and showcase the practicality of various potential techniques for 5G. New York University established NYU Wireless in August 2012 as a multi-disciplinary research center with a specific focus on advancing theories and foundational work for 5G wireless communications. Among the primary areas of interest was mmWave communications, operating in high-frequency bands above 10 GHz. Noteworthy milestones achieved by NYU Wireless included conducting the world's first radio channel measurements that confirmed the potential of mmWave spectrum and demonstrated the safety of mmWave radiation for human exposure. Shortly thereafter, in just two months following the open of NYU Wireless, the University of Surrey in the United Kingdom announced the establishment of a new 5G research center. This center received joint funding and technical support from the British government and a consortium comprising key mobile operators and vendors such as Huawei, Samsung, Telefonica, Fujitsu, and Rohde & Schwarz.

In November 2012, a flagship research project named *Mobile and wireless communications Enablers for the Twenty-twenty Information Society (METIS)*, which was funded by European Commission, was initiated (Osseiran et al., 2014). This project successfully established a global consensus on the nature of 5G before international standardization efforts like ITU-R and 3GPP committed. METIS's findings predicted a significant expansion in the number of interconnected devices, beyond the traditional human-centric communication prevalent in previous generations. This paradigm, commonly known as the Internet of Things (IoT) today, was expected to enhance daily lives by improving efficiency, comfortability, and security. Projections suggested that by 2020, the total number of IoT-connected devices could reach 50 billion. The coexistence of human-centric and machine-oriented applications was expected to bring a wide variety of unprecedented communication characteristics and requirements.

Unlike previous cellular systems that focused primarily on human-centric communication services, 5G aimed to broaden the scope of mobile communications. It extends beyond humans and includes objects, expanding from consumer applications to vertical industries, encompassing both public and private networks. This expansion resulted in a significant increase in potential mobile subscriptions, connecting not only billions of people but also enabling connectivity among machines and things. The disruptive nature of 5G opened up a variety of applications, such as Industry 4.0, virtual reality, the IoT, and autonomous driving. In summary, new-generation IMT systems need to support innovative use cases that require extremely high data rates, a massive number of connected devices, and low-latency high-reliability applications. 5G is essential for fulfilling the diverse requirements of various applications, offering additional capabilities that go beyond 4G mobile cellular networks. While legacy networks can handle certain applications, others demand strict latency and reliability, especially in critical areas like healthcare, security, logistics, automotive applications, and mission-critical control. Additionally, 5G supports a wide range of data rates, from multiple gigabits per second to ensuring high availability and reliability at tens of megabits per second. It also enables scalable and flexible networks to accommodate numerous devices while minimizing complexity and extending battery life.

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5_9

The age of 5G arrived in April 2019 when South Korea's mobile operators (SK Telecom, LG U+, and KT) and U.S. Verizon engaged in a debate over being the world's first provider of 5G communication services. Since then, we have witnessed a significant global expansion of 5G infrastructure and a remarkable surge in 5G subscriptions in major countries. As of this writing, there were 229 operational 5G networks worldwide, and users had access to over 700 different 5G smartphone models. For instance, the coverage of 5G in China hit 90 percent with the installation of around 3 million 5G base stations. According to GSMA, by the end of 2022, there were already over one billion consumer connections, and this number is projected to reach around 1.5 billion in 2023 and two billion by the end of 2025. These statistics highlight the fastest generational roll-out of 5G, surpassing the deployment speed of both 3G and 4G networks.

For the first time in the history of cellular generations, the term 5G has gained tremendous popularity, extending beyond technology and economics to become a focal point of geopolitical tensions. The year 2020 was marked by the emergence of the COVID-19 pandemic, which resulted in a significant loss of human life globally and posed unparalleled challenges to various societal and economic endeavors. However, this public health crisis served as a stark reminder of the crucial role played by networks and digital infrastructure in sustaining society and maintaining connections among families. Particularly, the value of 5G services and applications became evident in facilitating remote surgical procedures, online education, remote work arrangements, autonomous vehicles, unmanned deliveries, robotic assistance, intelligent healthcare systems, and automated manufacturing processes.

9.2 ITU-R Process of IMT-2020

Starting in 2012, the ITU-R Working Party 5D laid the groundwork for the next generation of IMT systems, formally referred to as IMT-2020. In February 2013, the ITU-R WP5D initiated two study items, one item to discuss the IMT Vision for 2020 and another to analyze future technological trends for terrestrial IMT systems. IMT-2020 aims to support a wide range of usage cases and applications beyond its predecessors. As a result of these studies, some findings were incorporated into the ITU-R M.2083 recommendation (ITU-R M.2083, 2015), which was published in 2015 and commonly referred to as the IMT vision for 2020 and beyond. This document outlines the framework and overall objectives of IMT-2020, serving as the first step in defining the new advancements, which include the future roles of IMT, its potential societal benefits, market, user, and technology trends, as well as spectrum implications.

In ITU-R M.2083 (2015), as shown in Fig. 9.1, three usage scenarios were defined for both human-centric and machine-oriented communications, i.e.:

- **Enhanced Mobile Broadband (eMBB):** Mobile broadband addresses the human-centric use cases for access to multimedia content, services, cloud, and data. With the proliferation of smart devices (smartphones, tablets, and wearable electronics) and the rising demand for video streaming, the need for mobile broadband continues to grow, setting new requirements for what ITU-R calls enhanced mobile broadband. This usage scenario comes with new use cases and requirements for improved capabilities and an increasingly seamless user experience. eMBB covers various cases, including wide-area coverage and hotspots, which have different requirements. For the hotspot case, i.e., for an area with high user density, very high traffic capacity is needed, while the requirement for mobility is low and user data rate is higher than that of wide-area coverage. Seamless coverage and medium to high mobility are desired for the wide-area coverage case, with a substantially improved data rate than existing data rates. However, the data rate requirement may be relaxed compared to hotspots.
- **Ultra-Reliable Low-Latency Communications (URLLC):** This scenario aims to support both human-centric and critical machine-type communications. It is a disruptive promotion over the previous generations of cellular systems that focused merely on the services for mobile subscribers. It opens the possibility for offering mission-critical wireless applications such as automatic driving, vehicle-to-vehicle communication involving safety, wireless control of industrial manufacturing or production processes, remote medical surgery, distribution automation in a smart grid, and transportation safety. It is characterized by stringent requirements such as low latency, ultra-reliability, and availability.
- **Massive Machine-Type Communications (mMTC):** This scenario supports massive connectivity with a vast number of connected devices that typically have very sparse transmissions of delay-tolerant data. Such devices, e.g., remote sensors, actuators, and monitoring equipment, are required to be low cost and low power consumption, allowing for a very long battery life of up to several years due to the possibility of remote IoT deployment.

As a parallel activity of examining the IMT vision for 2020, another study item focusing on the analysis of future technological trends for terrestrial IMT systems produced a report ITU-R M.2320 (2014) – *Future technology trends of*

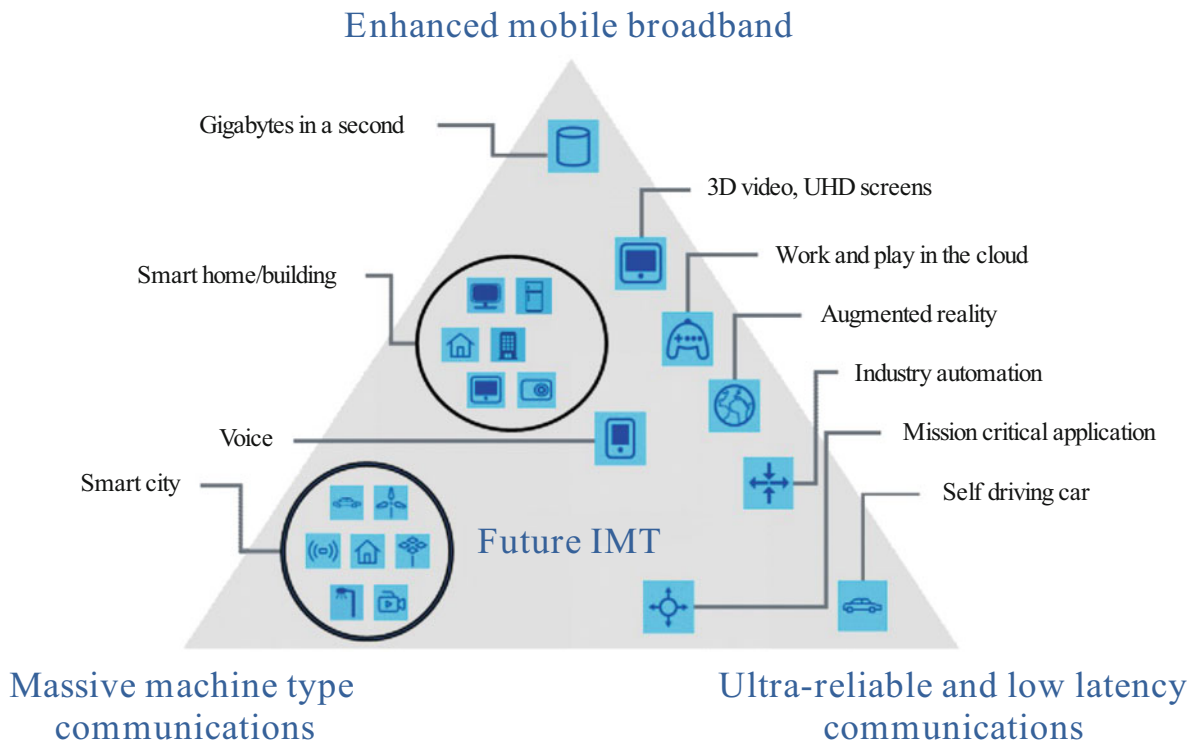


Fig. 9.1 Usage scenarios of IMT-2020. (Source: ITU-R M.2083, 2015)

terrestrial IMT systems. It provides detailed information about the technical and operational characteristics of IMT systems, including how they have evolved through advancements in technology and spectrally efficient techniques, as well as their deployment. Additionally, the report presents mobile traffic forecasts for the period leading up to 2015, sourced from various industry experts, and also includes a forecast for the years between 2015 and 2020, considering new market trends and drivers.

In order to support these market trends and to accommodate the mobile data traffic explosion, ITU-R M.2320 (2014) suggests considering the following aspects:

- *System average throughput*: The average throughput of cellular systems should be dramatically increased to support the exploding traffic by dramatically improving the spectrum efficiency.
- *User experience*: The user experience should be at least maintained regardless of the user's location and network traffic conditions.
- *Scalability*: The number of mobile terminals to be supported by a base station will be significantly increased due to the new services such as machine-to-machine (M2M) and IoT.
- *Latency*: Users' quality of experiences can be greatly improved by reducing the latency of the packet delivery and connection establishment.
- *Energy efficiency*: Low energy consumption is an important performance metric for both the network and the mobile devices.
- *Cost efficiency*: Low CAPEX and OPEX will reduce the cost of the network and may motivate operators to expand and improve their networks. Additionally, low-cost terminals will reduce the overall cost of a mobile subscription.
- *Network flexibility*: The ever-changing network topologies coupled with the complex and evolving wireless environment and services require that the future networks have a high degree of built-in flexibility to easily adapt to such changes as non-uniform traffic distribution in order to manage multiple generations of networks of different radio access technologies.
- *Non-traditional services*: Some potential new services and applications that are emerging in the mobile arena and are expected to undergo rapid development such as high-definition mobile video, M2M communication, enhanced location-based service, cloud computing, which bring new challenges in coverage, capacity, and user experience to a wireless network and consequently trigger the further improvement of wireless technologies.

- *Spectrum utilization*: More spectrum may be required to accommodate the mobile data traffic explosion. Many frequency arrangements, spanning a wide spectrum range, and increasing requirements to share with other services have resulted in multiple complex regulatory and technical considerations. While broad-based spectrum harmonization may reduce the cost of technology resources, addressing challenges such as shared use of spectrum, mobile network architecture optimization, RF component complexity, antenna efficiency, and device integration are the technology trends that have the potential to improve spectrum utilization.

Highlighting the development of technologies such as small cells, 3D beamforming, and massive MIMO that may realize their full potential when applied to smaller wavelengths, which are characteristic of higher frequency bands, extensive research has been conducted worldwide by different organizations to assess the feasibility of IMT in the spectrum exceeding 6 GHz. As a result, WP5D issued a dedicated report ITU-R M.2376 (2015) – *Technical feasibility of IMT in bands above 6 GHz*, aiming to investigate and provide insights into the practical viability of IMT within the frequency range of 6 GHz to 100 GHz. The technical feasibility encompasses an examination of current IMT systems, their advancements, potential new IMT radio interface technologies, and system approaches suitable for operation in the 6 GHz to 100 GHz bands. This evaluation took into consideration the influence of propagation characteristics associated with the potential future use of IMT in these frequency bands. Additionally, it considered technology enablers such as advancements in active and passive components, antenna techniques, and deployment architectures, as well as the outcomes of simulations and performance tests.

IMT-2020 was expected to provide far more enhanced capabilities than those of IMT-Advanced. In addition, IMT-2000 can be considered from multiple perspectives, including the users, manufacturers, application developers, network operators, and service and content providers. Therefore, it is recognized that technologies for IMT-2020 can be applied in a variety of deployment scenarios and can support a range of environments, service capabilities, and technology options (Andrews et al., 2014). Based on the usage scenarios and applications described as M.2083, the ITU-R defined a set of technical performance requirements. In November 2017, ITU-R released the recommendation – *Minimum requirements related to technical performance for IMT-2020 radio interface* (ITU-R M.2410, 2017), as the baseline for the evaluation of IMT-2020 candidate technologies. In addition to the peak data rates of 20 Gbps in the downlink and 10 Gbps in the uplink following the tradition of offering higher transmission rates, a number of other KPIs such as reliability, energy efficiency, and connection density were set up. The key capabilities of IMT-2020 are shown in Fig. 9.2, compared with those of IMT-Advanced. For providing readers with insight, Table 9.1 summarizes these performance requirements in a quantitative way. Note that different 5G use scenarios have significantly different profiles of performance requirements. More specifically, as illustrated in Fig. 9.3, while eMBB is demanding high experienced data rate up to Gbps and high area traffic capacity up to Tbps/km^2 level, mMTC focuses on high connection density that goes up to 1,000,000 devices per km^2, and URLLC calls for extremely high reliability (such as 99.9999%) within a strict latency constraint on ms level and zero mobility interruption time.

Another milestone for IMT-2020 development was the identification of its spectrum discussed in the ITU-R WRC. At WRC-15, a new set of frequency bands below 6 GHz (e.g., 470-694 MHz, 694-790 MHz, and 3300-3400 MHz) were identified for IMT on a global basis. This conference also appointed an agenda item for the following WRC-19 toward the identification of higher spectrum above 24 GHz for IMT-2020 mobile services. Based on the studies conducted by the ITU-R after WRC-15, WRC-19 noted that the ultra-low-latency and very high data rate applications require larger contiguous blocks of spectrum. As a consequence, a total of 13.5 GHz spectrum consisting of a set of high-frequency bands was assigned for the deployment of 5G mmWave communications:

- 24.25–27.5 GHz
- 37–43.5 GHz
- 45.5–47 GHz
- 47.2–48.2 GHz
- 66–71 GHz

In addition to ITU-R M.2410 (2017), the ITU-R also published other two documents in 2017: ITU-R M.2411 (2017) – *Requirements, evaluation criteria, and submission templates for the development of IMT-2020*, and ITU-R M.2412 (2017) – *Guidelines for evaluation of radio interface technologies for IMT-2020*. ITU-R M.2411 gives the submission and evaluation process for IMT-2020, which was initiated by Circular Letter 5/LCCE/59 and its Addenda. This document addresses the requirements, evaluation criteria, and submission templates necessary for a comprehensive submission of radio interface technologies for IMT-2020. On the other hand, ITU-R M.2412 outlines the guidelines for assessing candidate radio interface technologies for IMT-2020 in various test environments. These guidelines encompass the procedure, methodology,

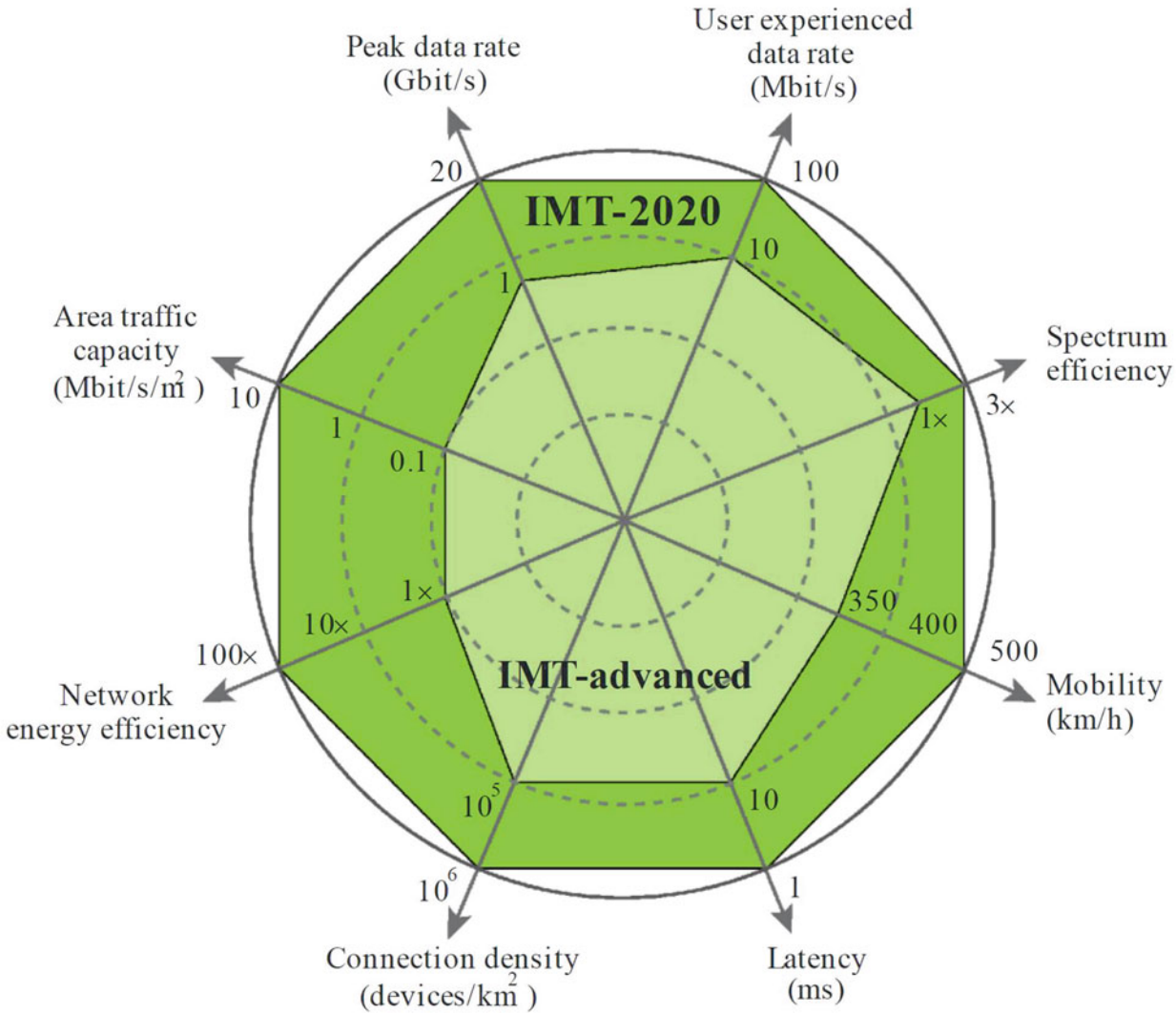


Fig. 9.2 Enhancement of key performance indicators from IMT-Advanced to IMT-2020. Source: (ITU-R M.2083, 2015)

and criteria (including technical, spectrum, and service aspects) to be employed during the evaluation process. The selected test environments closely replicate demanding radio operating conditions. Some of ITU-R technical reports and recommendations related to IMT-2020 are listed in Table 9.2.

The evaluation and submission processes for IMT-2020 began in March 2016 when the ITU-R issued an invitation for proposals for candidate Radio Interface Technologiess (RITs) or Sets of Radio Interface Technologiess (SRITs) for the terrestrial components of IMT-2020 as the development process illustrated in Fig. 9.4. By the end of 2019, a total of seven candidate IMT-2020 proposals had been submitted, and evaluation reports from independent evaluation groups had been received. Among these submissions were two independent proposals from 3GPP for IMT-2020, which were not interconnected. The first proposal was NR, submitted as a RIT proposal for IMT-2020. In 3GPP, 5G NR is the name given to the 5G radio interface. The second proposal consisted of NR and E-UTRA jointly submitted as two-component RITs within a set of radio interfaces. China and South Korea submitted two proposals that were identical to the RIT submissions made by 3GPP. The remaining three proposals were ETSI/DECT-2020, Nufront Enhanced Ultra High-Throughput (EUHT), and the 5Gi standard proposed by *Telecommunications Standards Development Society, India* (TSDSI).

The evaluation of candidate IMT-2020 radio interfaces ended officially in February 2021, resulting in the approval of two 3GPP technologies (5G-SRIT and 5G-RIT) and 5Gi. The 5Gi technology is primarily based on 3GPP NR, with the introduction of an additional component known as LMLC (Low Mobility Low Cost) to enable affordable 5G connectivity in rural areas. These results are incorporated in the recommendation ITU-R M.2150 (2022), which outlines the detailed specifications for the IMT-2020 radio interfaces, namely 3GPP SRIT, 3GPP RIT, and TSDSI 5Gi.

Table 9.1 Minimum Technical Performance Requirements for IMT-2020 (Jiang & Luo, 2023)

KPI	Minimum performance requirement
Peak Data Rate	Downlink: 20 Gbps
	Uplink: 10 Gbps
Peak Spectral Efficiency	Downlink: 30 bps/Hz
	Uplink: 15 bps/Hz
User-Experienced Rate	Downlink: 100 Mbps
	Uplink: 50 Mbps
5th-Percentile User Spectral Efficiency	Downlink: 0.12 bps/Hz∼0.3 bps/Hz
	Uplink: 0.045 bps/Hz∼ 0.21 bps/Hz
Average Spectral Efficiency	Downlink: 3.3 bps/Hz∼ 9 bps/Hz
	Uplink: 1.6 bps/Hz∼6.75 bps/Hz
Area Traffic Capacity	10 Mbps/m^2 (Indoor Hotspot)
User Plane Latency	4 ms—eMBB
	1 ms—URLLC
Control Plane Latency	20 ms
Connection Density	1,000,000 devices per km^2
Energy Efficiency	The support for two aspects:
	(1) Efficient data transmission in a loaded case
	(2) Low energy consumption when there are no data
Reliability	1–10^{-5} (99.999%)
Mobility	Up to 500 km/h
Mobility Interruption Time	0 ms
Maximal Bandwidth	100 MHz for sub-6 GHz
	1 GHz for mmWave

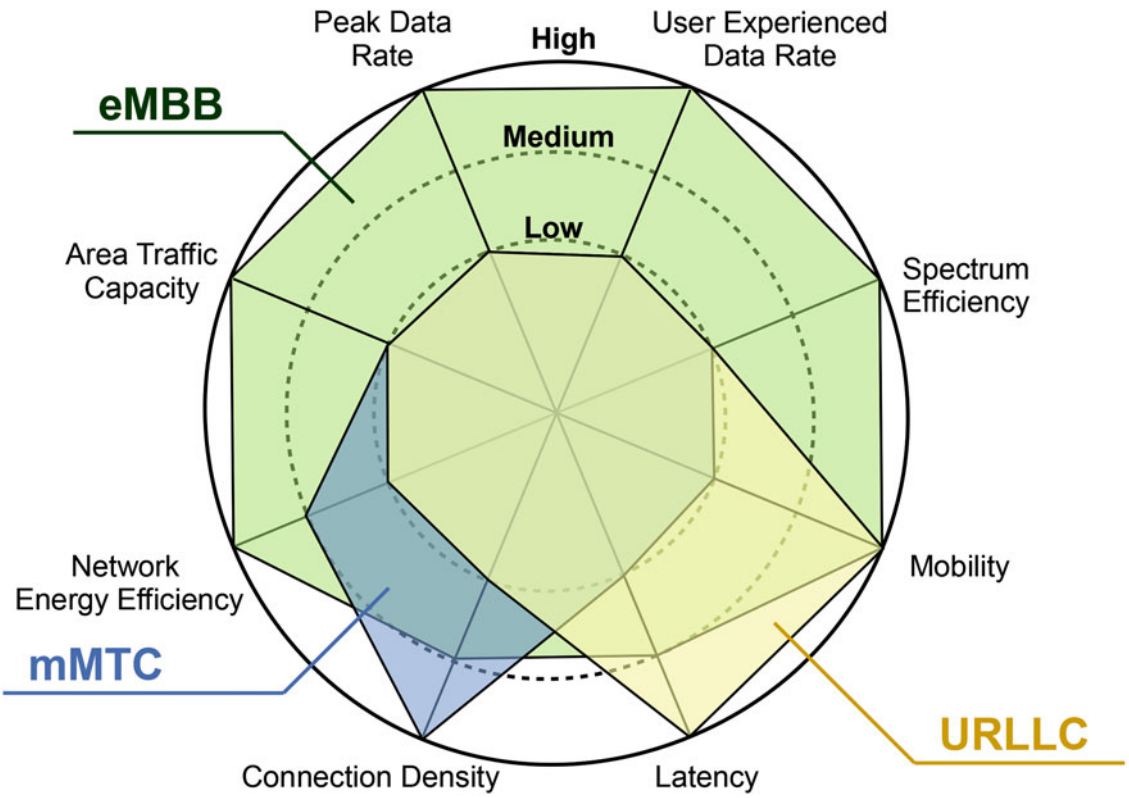


Fig. 9.3 The key capabilities in different 5G use scenarios

Table 9.2 Major ITU-R Reports and Recommendations for IMT-2020

Document	Time	Main content
M.2083	Sep. 2015	IMT vision framework and overall objectives of the future development of IMT for 2020 and beyond
M.2320	Nov. 2014	Future technology trends of terrestrial IMT systems
M.2376	Jul. 2015	Technical feasibility of IMT in bands above 6 GHz
M.2410	Nov. 2017	Minimum requirements related to technical performance for IMT-2020 radio interface(s)
M.2411	Nov. 2017	Requirements, evaluation criteria, and submission templates for the development of IMT-2020
M.2412	Oct. 2017	Guidelines for evaluation of radio interface technologies for IMT-2020
M.2150	Feb. 2022	Detailed specifications of the terrestrial radio interfaces of International Mobile Telecommunications-2020 (IMT-2020)

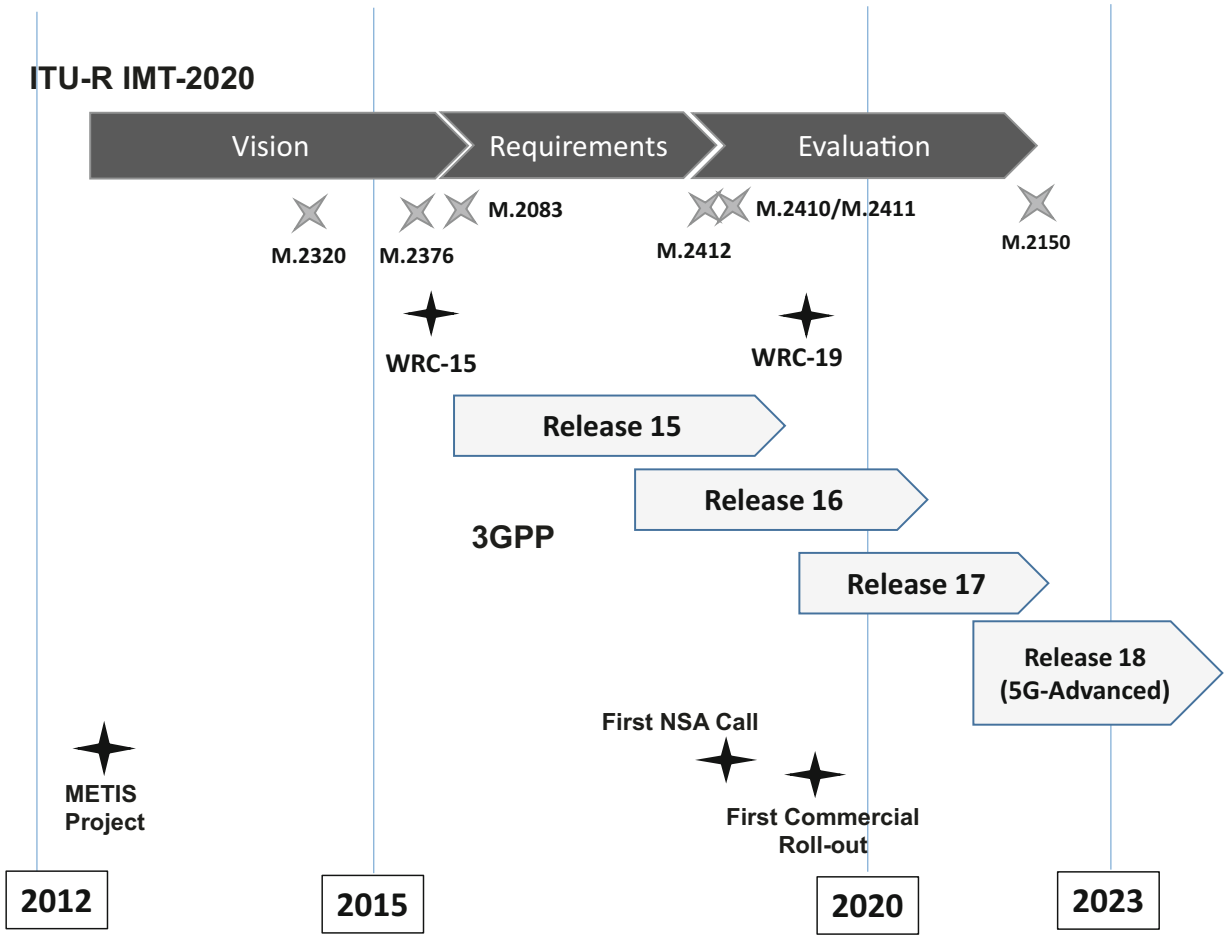


Fig. 9.4 The process for defining and standardizing 5G/IMT-2020

9.3 3GPP Standardization for 5G

With the IMT-2020 framework defined by the ITU-R and the frequency bands identified by the WRC, the task of specifying detailed technologies fell into the standardization bodies. In contrast to previous generations, which involved conflicting technical paths and multiple standardization authorities, the development of 5G standards was primarily driven by the dominant role of 3GPP. The technical specifications organized as releases by 3GPP actually served as the de facto standards.

As early as 2015, the 3GPP RAN group made the decision to establish a study item in Release 14 focused on 5G New Radio (NR) and began the task of channel modeling for frequency bands above 6 GHz. The specification of the initial NR took place as part of a work item in Release 15. However, in order to meet commercial demands for early large-scale trials and deployments in 2018, which was earlier than the originally anticipated timeline of around 2020, 3GPP committed to

expediting the process. They agreed to prioritize the finalization of a NSA variant. By the end of 2017, the first version of the 5G specifications became available. In an NSA deployment, the NR air interface is connected to the existing EPC core network, allowing the capabilities offered by NR (such as lower latency) to be utilized without the need for network replacement.

Just before the Mobile World Congress commenced on February 26, 2018, the world's first 5G NSA call was successfully completed in Spain through a collaborative effort between Vodafone and Huawei. Following the initial delivery of the NSA variant, a significant amount of effort by 3GPP was redirected toward ensuring the timely completion of Release 15, which aimed to establish the first comprehensive set of 5G standards. In parallel to the NR radio access technology, 3GPP also developed a new core network known as 5GC. In June 2018, the final version of Release 15, capable of supporting the Standalone (SA) operation of 5G NR, became available. This marked the completion of 5G Phase 1. The focus of Release 15 primarily revolved around eMBB and, to some extent, URLLC. Meanwhile, mMTC continued to be supported through LTE-based machine-type communication technologies such as eMTC and NB-IoT. Release 15 provided the foundation on which 3GPP continues its work to evolve the capability and functionality of 5G so as to support new spectrum and new applications and further enhance existing core features.

Compared with its predecessor, i.e., LTE-Advanced, 5G NR offers many technical advantages (Dahlman et al., 2021):

- To achieve exceptionally high data rates, the utilization of millimeter-wave frequency bands is employed to acquire abundant spectral resources and wide transmission bandwidth. NR is capable of operating in a frequency band over 6 GHz. This allows for heterogeneous deployment, wherein macro base stations operate at lower carrier frequencies, while small base stations operate at higher carrier frequencies.
- By adopting an ultra-lean design approach, network energy efficiency can be effectively improved, and interference can be reduced, especially during high traffic load conditions. Synchronization signals, system broadcast information, and reference signals are only transmitted when necessary, rather than continuous *always on* transmissions employed in previous generations.
- Forward compatibility is implemented to proactively accommodate future enhancements for supporting yet unknown use cases and applications. This is accomplished through the principles of *self-containment* and *well-confined transmission*. Self-containment ensures that data within a specific slot and beam can be detected independently without reliance on other slots or beams. Well-confined transmission focuses on confining transmissions within specific frequency and time domains, thereby enabling the incorporation of new types of transmissions alongside existing legacy transmissions in the future. It allows for seamless integration of future advancements without disrupting the existing system (Zaidi et al., 2017).
- System flexibility adapts to a wide range of carrier frequencies, heterogeneous deployment (macro-, micro-, and pico-cells), and diverse usage scenarios (eMBB, URLLC, and mMTC) with stringent and sometimes contradictory requirements. The physical layer design of NR is flexible and scalable, including highly adaptive modulation schemes (from BPSK in the uplink of mMTC to 1024QAM in the downlink of eMBB), a scalable OFDM numerology, LDPC codes with rate-compatible structure, and a flexible frame structure.
- Beam-centric design to enable the extensive usage of massive MIMO for not only data transmission but also control signaling. Reference signals can be beam formed with configurable granularity in the time domain and frequency domain.

In contrast to previous generations, the 3GPP 5G system adopted the service-based architecture (SBA) that organizes the architectural elements into a set of service-oriented network functions. The interaction between network functions is represented in two ways:

- *Service-based representation*, where network functions enable other authorized network functions to access their services via the interfaces of a common framework
- *Reference point representation*, where the interaction between any two network functions is described by a point-to-point reference point

The 5G System (5GS) architecture is designed to support a wide variety of use cases with stringent but sometimes conflicting performance requirements. It enables deployments to use techniques such as NFV, SDN, and network slicing. Some key principles and concept (3GPP TS 23.501, 2021) are to:

- Separate the user plane functions from the control plane functions, allowing independent scalability, evolution, and flexible deployments.
- Modularize the function design to enable flexible and efficient network slicing. Wherever applicable, defining procedures (i.e., the set of interactions among network functions) as services, to maximize the re-usability.

- Enable each network function and its network function services to interact with other network functions directly or indirectly via a service communication proxy if required.
- Minimize dependencies between the access network and the core network. The architecture is defined as a converged core network with a common interface integrating both 3GPP and non-3GPP access technologies.
- Support a unified authentication framework.
- Support stateless network functions, where the computing resource is decoupled from the storage resource.
- Support capability exposure.
- Support concurrent access to local and centralized services. To support low-latency services and access to local data networks, UP functions can be deployed close to the access network.
- Support roaming with both home-routed traffic and local breakout traffic in the visited PLMN.

While there are similarities in certain functionalities between the 5G core network and previous generations, this resemblance is expected because fundamental operations, such as authentication, charging, resource allocation, and mobility management, are inherent to any network. However, noteworthy new functions have emerged to support novel network paradigms, including network slicing, service-based networking, and user/control plane separation. To accommodate diverse data services and varying requirements, the network functions have undergone further simplification. Many of these functions have transitioned to software-based implementations, enabling them to run on generic computer hardware. This software-based approach enhances flexibility and facilitates the enablement of different services within the network.

The evolution of 5G NR continued in Release 16, often referred to informally as "5G Phase 2," which was completed in June 2020. This release aimed to fulfill the IMT-2020 requirements and, alongside Release 15, served as the initial comprehensive set of 3GPP 5G specifications submitted to the ITU-R. The detailed requirements of some typical 5G use cases, as per definition in Release 16 (3GPP TS 22.261: Service Requirements for the 5G System; Stage 1; V16.16.0, 2021), are listed in Tables 9.3 and 9.4, respectively. The final proposal from 3GPP consisted of two distinct submissions known as the single RIT and the combined SRIT. In November 2020, the ITU-R announced that both the 3GPP SRIT and RIT adhered to the vision and rigorous performance requirements outlined by the IMT-2020 framework. This endorsement confirmed the alignment of 3GPP's 5G specifications with the IMT-2020 vision and standards.

Release 16 introduced significant enhancements for both the NR and LTE radio interfaces, focusing on improving their collaboration. Within the 5G NR, user data rates have been increased through various carrier aggregation configurations and the use of 256QAM modulation. Additional enhancements include NR-based access to unlicensed spectrum, improvements in mobility, and enhancements in UE power saving. The overarching goal of Release 16 is to transform the 5GS into a versatile communication platform capable of serving a wide range of industries or verticals (Baek et al., 2021), such as transportation (autonomous driving, V2V, railways, maritime), automated factories, healthcare, public safety, and more. To achieve this, the 5GS's versatility and reliability have been further enhanced to meet industry-grade requirements. Notable enhancements include improvements in URLLC, network slicing, edge computing, cellular IoT, non-public networks, positioning services, and LAN-type services. In addition, the utilization of 5G as an underlying communication network, transparent to external applications, has been improved through the introduction of *Northbound Application Programming Interface (API)*. Beyond industrial applications, Release 16 encompasses enhancements for the coexistence of 5G with non-3GPP systems, entertainment services such as streaming and media distribution, and network optimizations such as user identity management.

Release 17, the third version of 5G specifications, was frozen in mid-2022. Notably, this was the first and only release of 3GPP that was completed entirely remotely, using email discussions and online sessions, due to the travel restrictions imposed during the COVID-19 pandemic. It is probably the most versatile release in the history of 3GPP in terms of the number of technical features, as shown in Fig. 9.5. Release 17 focuses on consolidating and enhancing the concepts and functionalities introduced in previous releases, while also introducing a limited number of new features. These enhancements cover various key areas addressed in earlier releases. First, improvements are made to services for different verticals, including advancements in positioning, private networks, and more. Second, enhancements are introduced for 5G support of IoT applications, in both the core network and access network, as well as proximity-based (direct) communications between mobile devices, particularly in the context of autonomous driving. Additionally, enhancements target different aspects of media-related user plane services in the entertainment industry, such as codecs, streaming, and broadcasting. Mission-critical communications are also addressed within Release 17. Furthermore, several network functionalities have been improved, including slicing, traffic steering, and edge computing. The radio interface and access network have undergone significant improvements as well, such as advancements in MIMO technology, relays, and the use of 1024QAM modulation for the downlink.

Table 9.3 5G performance requirements for high data rate and traffic density scenarios defined in 3GPP Release 16 (3GPP TS 22.261: Service Requirements for the 5G System; Stage 1; V16.16.0, 2021)

	Scenario	Experi-enced data rate DL	Experi-enced data rate UL	Area traffic capacity DL	Area traffic capacity UL	Overall user density	Activity factor	UE speed	Coverage
1	Urban macro	50 Mbps	25 Mbps	100 Gbps/km^2 (note 4)	50 Gbps/km^2 (note 4)	10000/km^2	20%	Pedestrians and users in vehicles (up to 120 km/h)	Full network (note 1)
2	Rural macro	50 Mbps	25 Mbps	1 Gbps/km^2 (note 4)	500 Mbps/km^2 (note 4)	100/km^2	20%	Pedestrians and users in vehicles (up to 120 km/h)	Full network (note 1)
3	Indoor hotspot	1 Gbps	500 Mbps	15 Tbps/km^2	2 Tbps/km^2	250000/km^2	note 2	Pedestrians	Office and residential (note 2) (note 3)
4	Broadband access in a crowd	25 Mbps	50 Mbps	3.75 Tbps/km^2	7.5 Tbps/km^2	500000/km^2	30%	Pedestrians	Confined area
5	Dense urban	300 Mbps	50 Mbps	750 Gbps/km^2 (note 4)	125 Gbps/km^2 (note 4)	25000/km^2	10%	Pedestrians and users in vehicles (up to 60 km/h)	Downtown (note 1)
6	Broadcast-like services	Maximum 200 Mbps per TV channel	N/A or modest (e.g., 500 kbps per user)	N/A	N/A	15 TV channels of 20 Mbps on one carrier	N/A	Stationary users, pedestrians, and users in vehicles (up to 500 km/h)	Full network (note 1)
7	High speed train	50 Mbps	25 Mbps	15 Gbps/train	7.5 Gbps/train	1000/train	30%	Users in trains (up to 500 km/h)	Along railways (note 1)
8	High speed vehicle	50 Mbps	25 Mbps	100 Gbps/km^2	50 Gbps/km^2	4000/km^2	50%	Users in vehicles (up to 250 km/h)	Along roads (note 1)
9	Airplanes connectivity	15 Mbps	7.5 Mbps	1.2 Gbps/plane	600 Mbps/plane	400/plane	20%	Users in airplanes (up to 1000 km/h)	(note 1)

Note 1: For users in vehicles, the UE can be connected to the network directly, or via an on-board moving base station
Note 2: A certain traffic mix is assumed; only some users use services that require the highest data rates (Alliance, 2015)
Note 3: For interactive audio and video services, for example, virtual meetings, the required two-way end-to-end (E2E) latency (UL and DL) is 2–4 ms, while the corresponding experienced data rate needs to be up to 8K three-dimensional (3D) video (300 Mbps) in UL and DL
Note 4: These values are derived based on overall user density. Detailed information can be found in (Alliance, 2016b)
Note 5: All the values in this table are targeted values and not strict requirements

3GPP has introduced the official new name of *5G-Advanced* to signify the evolution of 5G in Release 18 and beyond (Chen et al., 2022). As of this writing, the specification of Release 18 is underway within 3GPP, with an anticipated completion date by the end of 2023 as per the current schedule. A key focus of Release 18 is the further improvement of energy efficiency in both air interface technology and network infrastructure. Efforts are being made to identify additional power-saving measures and explore possibilities for cost reduction for Reduced Capability (RedCap) devices, which were defined in Release 17. One area of investigation is the potential use of energy harvesting technologies. The term "National Security and Public Safety" (NSPS) refers to functionalities that ensure reliable and secure communication, particularly for emergency services and rescue operations. Release 18 also aims to support various forms of augmented reality and virtual reality services through the XR (Extended Reality) concept. Moreover, the integration of artificial intelligence and machine learning techniques is planned to optimize the physical layer and enhance transmission and reception technologies. Lastly, Release 18 explores the use of full-duplex data transmission at base stations, allowing for simultaneous transmission and reception via TDD bands.

Table 9.4 5G performance requirements for low-latency and high-reliability scenarios defined in 3GPP Release16 (3GPP TS 22.261: Service Requirements for the 5G System; Stage 1; V16.16.0, 2021)

	Scenario	Max. E2E latency (note 2)	Survival time	Comm. service availa-bility (note 3)	Relia-bility (note 3)	User-experienced data rate	Payload size (note 4)	Traffic per km^2 (note 5)	Connec-tions per km^2 (note 6)	Service area dimension (note 7)
1	Discrete automation	10 ms	0 ms	99.99%	99.99%	10 Mbps	Small to big	1 Tbps	100000	1000×1000 ×30 m
2	Process automation: remote control	60 ms	100 ms	99.9999%	99.999%	1 to 100 Mbps	Small to big	100 Gbps	1000	300×300 ×50 m
3	Process automation: monitoring	60 ms	100 ms	99.9%	99.9%	1 Mbps	Small	10 Gbps	10000	300×300 ×50 m
4	Electricity distribution: medium voltage	40 ms	25 ms	99.9%	99.9%	10 Mbps	Small to big	10 Gbps	1000	100 km along power line
5	Electricity distribution: high voltage (note 1)	5 ms	10 ms	99.9999%	99.999%	10 Mbps	Small	100 Gbps	1000 (note 8)	200 km along power line
6	Intelligent transport systems: infrastructure backhaul	30 ms	100 ms	99.9999%	99.999%	10 Mbps	Small to big	10 Gbps	1000	2 km along road

Note 1: Currently realized via wired communication lines
Note 2: This is the maximum E2E latency allowed for the 5G system to deliver the service in the case the E2E latency is completely allocated to the 5G system from the UE to the Interface to Data Network
Note 3: Communication service availability relates to the service interfaces, and reliability relates to a given system entity. One or more retransmissions of network layer packets may take place in order to satisfy the reliability requirement
Note 4: Small: payload typically $\leqslant$ 256 bytes
Note 5: Based on the assumption that all connected applications within the service volume require the user-experienced data rate
Note 6: Under the assumption of 100% 5G penetration
Note 7: Estimates of maximum dimensions; the last figure is the vertical dimension
Note 8: In dense urban areas
Note 9: All the values in this table are example values and not strict requirements. Deployment configurations should be taken into account when considering service offerings that meet the targets

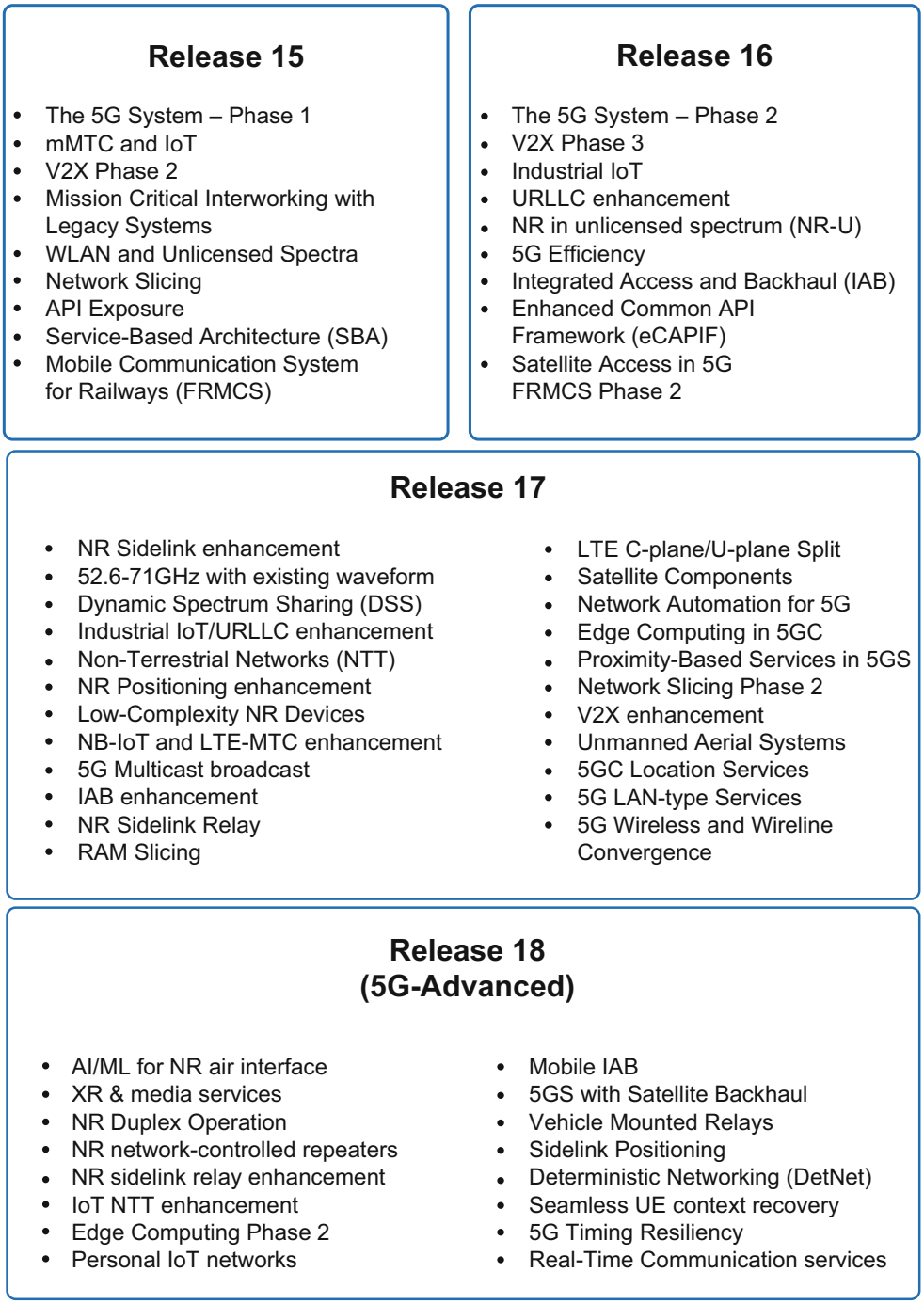


Fig. 9.5 Main technical features of 3GPP 5G releases

9.4 5G Key Technologies

To fulfill the stringent performance requirements for IMT-2020, as defined in ITU-R M.2410 (2017), the 5G system adopted disruptive radio access and networking technologies. While the 5G system integrated numerous techniques, the key advancements that stand out are massive MIMO, mmWave, NOMA, polar codes, network function virtualization, software-defined networking, and network slicing. This section aims to offer readers a comprehensive understanding of these key 5G technologies, including their principles, benefits, issues, and obstacles.

9.4.1 Massive MIMO

From the perspective of a cellular network, a base station must be capable of supporting multiple active devices simultaneously. Single-user multiple-input multiple-output (SU-MIMO) is a system that uses multiple transmit and receive antennas to deliver data to a single user, enabling spatial multiplexing. However, it does not mean that the whole system is limited to a single user. Instead, a large number of users can be accommodated by using separate time–frequency resources through *time-division multiplexing* and *frequency-division multiplexing*.

In theory, the channel capacity of SU-MIMO increases linearly as the number of transmit and receive antennas increases. However, there are three practical limitations that prevent SU-MIMO from scaling to high-order spatial multiplexing. First, it is difficult to incorporate a large number of antennas into a user terminal and employ advanced signal processing algorithms to separate high-dimensional data streams due to hardware size, power supply, and cost constraints. Second, even in environments with rich scattering, a compact antenna array struggles to support numerous independent sub-channels in a point-to-point connection. This is particularly true in line-of-sight conditions where the channel matrix has a minimal rank of one. Third, the channel capacity experiences slow growth at low SNR, such as at the edge of a cell where most devices are typically located. Cell-edge users suffer from high path loss and strong interference from neighboring cells, limiting the use of capacity-approaching techniques (Marzetta, 2015).

Multi-user multiple-input multiple-output (MU-MIMO) surpasses SU-MIMO by breaking up spatially multiplexed data streams among multiple terminals, offering two primary advantages. First, MU-MIMO only necessitates single-antenna terminals, enabling the use of affordable, low-complexity, and power-efficient equipment. Second, it exhibits greater resilience to propagation conditions, relying on the spatial arrangement of terminals. It can perform effectively even in line-of-sight scenarios, as long as the angular separation between terminals exceeds the angular resolution of the base station array. However, conventional MU-MIMO still faces challenges when scaling up for high-order spatial multiplexing, primarily due to the exponentially increasing complexity of capacity-achieving precoding and decoding. The most significant obstacle is that the transmitter needs knowledge of the downlink channel, and the resources required for acquiring CSI escalate with the number of service antennas and users.

Massive MIMO, also known as large-scale antenna array, very large MIMO, hyper MIMO, and full-dimension MIMO, breaks this scalability barrier by not attempting to achieve the full Shannon limit and paradoxically by increasing the size of the system. It departs from Shannon-theoretic practice in three ways (Marzetta, 2010):

- The base station is the only entity that acquires channel knowledge for precoding in the downlink and detection in the uplink, while the terminals do not use this information. Leveraging the channel reciprocity in a TDD system, the overhead for acquiring CSI depends on the number of terminals, regardless of the massive number of antennas at the base station.
- The base station deploys a large-scale antenna array, allowing the number of service antennas to be several times greater than the number of active users. By keeping the number of users relatively small, the implementation complexity of the system remains moderate.
- In the downlink, linear precoding such as conjugate beamforming and zero-forcing precoding (Jiang & Schotten, 2022a) is utilized, accompanied by linear detection in the uplink. As the number of base station antennas rises, the performance of linear precoding and detection approaches the Shannon limit.

By employing a substantial number of antennas, the transmission energy can be precisely focused on a highly localized region. This directivity yields significant enhancements in spectral efficiency and energy efficiency. Moreover, massive MIMO achieves cost efficiency by leveraging affordable, low-precision RF components, eliminating the need for expensive, high-linearity power amplifiers used in traditional systems. Instead, it utilizes numerous inexpensive power amplifiers with output power in the milli-Watt range. In addition to cost advantages, massive MIMO offers several other benefits, including reduced air interface latency, simplified multiple access layer, and increased resilience against both unintentional interference and deliberate jamming (Larsson et al., 2014).

At low frequencies, NR employs a low to moderate number of antennas, typically up to 64 transmit and receive antennas at the base station side operating around 700 MHz (3GPP TR38.913, 2020). In this scenario, FDD operation is supported, requiring the transmission of a CSI-RS in the downlink for accurate CSI acquisition and CSI reporting in the uplink. Due to the limited available bandwidth in this frequency range, achieving high spectral efficiency is crucial, which is facilitated by MU-MIMO with high-order spatial multiplexing. Compared to LTE, NR utilizes high-resolution CSI reporting to enhance this efficiency.

For higher frequencies, a larger number of antennas can be employed within the same hardware size, with NR supporting up to 256 transmit and receive antennas around 4 GHz. The large antenna array enables advanced beamforming and MU-

MIMO capabilities. As the number of reference signals is proportional to the number of transmit antennas, massive MIMO is preferred to operate in the TDD mode taking advantage of channel reciprocity. In this configuration, the base station acquires downlink CSI by estimating the channel-sounding reference signals transmitted in the uplink. During downlink data transmission, there are no reference signals, and linear precoding schemes such as conjugate beamforming and zero-forcing precoding are applied to simplify signal detection at the UE side.

At even higher frequencies, specifically in the mmWave range, the implementation typically requires analog beamforming or hybrid beamforming (Jiang & Schotten, 2022c). This limitation restricts transmission to a single beam direction per time unit and radio chain. Due to the short carrier wavelength in this frequency range, the size of an isotropic antenna element becomes small, enabling to use a large number of antenna elements to maintain coverage. Beamforming becomes essential for both the transmitter and receiver to combat the increased path loss, even for control channel transmission. A new beam management process is required for CSI acquisition, involving sequential sweeping of radio transmitter beam candidates by the base station over time, while the UE maintains an appropriate radio receiver beam to receive the selected transmitter beam. In the frequency bands around 30 GHz and 70 GHz, NR supports up to 256 antenna elements at the base station side and up to 32 antenna elements at the UE side.

9.4.2 Non-orthogonal Multiple Access/NOMA

Multiple access is a critical technique in cellular communication systems that allows multiple users to share radio resources. Over the years, there have been significant advancements in the multiple access schemes employed by cellular systems. These schemes, including FDMA, TDMA, CDMA, and OFDMA, have been implemented in 1G to 4G systems, respectively. They fall under the category of OMA, where each user is assigned to an orthogonal radio resource unit in terms of frequency, time, code, or a combination of these domains. OMA has been widely used in previous generations of cellular communications because it simplifies transceiver design and reduces interference among users. Despite the advantage of low-complexity multi-user detection at the receiver, it is widely acknowledged that OMA falls short of achieving the maximum sum-rate capacity in multi-user wireless systems (Chen et al., 2018).

To fulfill the diverse requirements of massive connectivity, high spectral efficiency, low latency, and fairness, the 5G system has considered a novel technique known as non-orthogonal multiple access (NOMA) as one of its multiple access methods. Unlike traditional OMA schemes, NOMA stands out by accommodating a greater number of users than the available orthogonal resource units through non-orthogonal resource sharing (Jiang & Schotten, 2023a). To realize NOMA, a technique called superposition coding at the transmitter and SIC at the receiver have been introduced. This approach allows multiple users to reuse each orthogonal resource unit. In superposition coding, the individual information symbols from different users are combined into a single waveform at the transmitter. On the receiver side, SIC is employed to iteratively decode the signals, progressively eliminating interference until the desired signal is obtained. This approach is sometimes referred to as power-domain NOMA.

In the case of two users, as illustrated in Fig. 9.6, a far user at the cell edge is allocated to more power and a near user at the cell center gets less power. Regardless of the difference of channel gains, this power ratio will be kept in the received signal of any user. The near user first detects the symbol of the far user and then subtracts its regenerated component from the received signal to detect its own symbol. The far user detects its signal directly by treating the signal of the near user as a colored noise.

A study by Ding et al. (2017) provides a simple example to illustrate the advantages of NOMA over OMA. Consider a scenario where a user with a very weak channel condition requires fair treatment, due to either high-priority data or a prolonged lack of service. In OMA, one of the scarce bandwidth resources must be exclusively allocated to this user, despite their poor channel conditions. This design negatively impacts spectrum efficiency and overall system capacity. In contrast, NOMA ensures that the user with poor channel conditions is served while allowing users with better channel conditions to concurrently access the same resources. As a result, NOMA achieves user fairness and can effectively boost system capacity compared to OMA. Apart from its spectral efficiency gains, NOMA is able to support a larger number of users, facilitating massive connectivity in IoT scenarios.

LTE Release 13 discussed Multi-user Superposed Transmission (MUST) as a specific instance of the NOMA technique, primarily focused on the downlink transmission (Yuan et al., 2015). In the MUST schemes, adaptive power control and bit labeling are employed at the transmitter side, categorizing them into three groups. Category 1 maps coded bits of co-scheduled users to component constellation symbols without Gray mapping independently. Category 2 jointly maps coded bits of co-scheduled users to component constellations using Gray mapping. Category 3 directly maps coded bits onto symbols of a composite constellation. During the study item in Release 14, various NOMA schemes such as Sparse Code

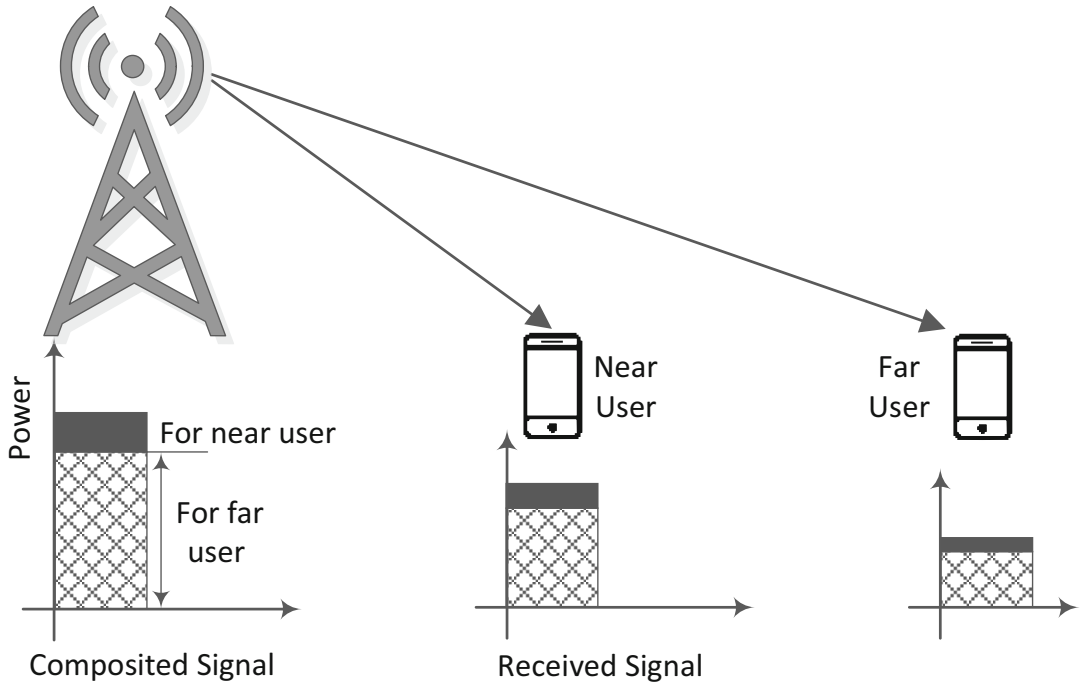


Fig. 9.6 Schematic diagram of downlink power-domain NOMA

Multiple Access (SCMA), Multi-user Shared Access (MUSA), Pattern Division Multiple Access (PDMA), and Resource Spread Multiple Access (RSMA) were proposed. Release 14 LTE specified grant-based NOMA for downlink eMBB support, typically operating in the RRC connected state. In 3GPP Release 15, a study item was established to further investigate signal processing at the transmitter side, multi-user receiver design, complexity analysis, and NOMA-related procedures such as HARQ, link adaptation, and power allocation. However, since massive MIMO brought significant performance gains in NR, applying NOMA in the NR downlink only provides marginal improvements. Consequently, the focus of the NOMA study item in Release 16 shifted to uplink grant-free transmission, aiming to reduce control signaling overhead, transmission latency, and device power consumption.

9.4.3 Millimeter-Wave/mmWave Communications

Previous generations of cellular systems definitely operated within low-frequency bands spanning from several hundred megahertz to several gigahertz. On one hand, the availability of spectral resources in these bands is quite limited compared to the increasing demand for mobile broadband services. On the other hand, these bands are utilized by a diverse range of applications such as television broadcasting, satellite communications, radar, radio astronomy, aviation, and maritime navigation, leading to a global shortage of spectrum. The scarcity of resources has prompted mobile network operators to investigate the under-utilized mmWave spectrum as a means to offer mobile broadband services. The mmWave spectrum, also referred to as the millimeter band, encompasses electromagnetic wavelengths ranging from 1 mm (corresponding to 300 GHz) to 10 mm (equivalent to 30 GHz). It was envisioned that mmWave could bring about several advantages in mobile communications, including cost-effective fiber replacements for mobile backhaul, dense deployment of mmWave small cells, wireless broadband access, and low-latency delivery of uncompressed high-definition media (Rappaport et al., 2013).

Extensive research has been conducted worldwide by different organizations to assess the feasibility of IMT in the spectrum exceeding 6 GHz. As a result, ITU-R WP5D issued a dedicated report ITU-R M.2376 (2015) – *Technical feasibility of IMT in bands above 6 GHz*, aiming to investigate and provide insights into the practical viability of IMT within the frequency range of 6 GHz to 100 GHz. WRC-15 appointed an agenda item for the following WRC-19 toward the identification of higher spectrum above 24 GHz for IMT-2020 mobile services. Based on the studies conducted by the ITU-R afterward, WRC-19 identified a total of 13.5 GHz bandwidth over the frequency range from 24.25 GHz to 71 GHz for the deployment of 5G mmWave communications. 3GPP defined the relevant spectrum for NR, which was divided into

two frequency ranges: the First Frequency Range (FR1), including sub-6 GHz frequency bands extending from 450 MHz to 6 GHz, and the Second Frequency Range (FR2) covering 24.25 GHz to 52.6 GHz (Dahlman et al., 2021).

MmWave wireless communications offer great promise due to the abundance of spectral resources available within the frequency range of 30 GHz to 300 GHz. This range provides a substantial amount of contiguous spectrum, which helps address the bandwidth limitations experienced at sub-6 GHz frequencies. However, the practical implementation of mmWave technology in mobile networks presents significant technical challenges when it comes to designing and developing RF and antenna components. One major challenge stems from atmospheric losses caused by the absorption of mmWave signals by water vapor and oxygen (Siles et al., 2015), surpassing the losses typically encountered in free space. Additionally, mmWave signals generally face difficulties in penetrating solid materials such as reinforced concrete walls, resulting in greater propagation and penetration losses compared to microwave signals.

To mitigate the considerable propagation losses, antenna arrays are essential at both the base station and UE sides. These arrays concentrate the transmission energy into a smaller region, compensating for the large power loss over mmWave signals. The compact size of mmWave transceiver antennas is a result of the short wavelength of high-frequency signals, ranging from 1 mm to 10 mm. This compactness enables the use of large-scale antenna arrays. However, the differences in antenna size and propagation characteristics impose various challenges and constraints on the design of physical layer transmission algorithms and medium-access protocols for mmWave systems. Unlike conventional cellular systems that provide seamless coverage for a large number of users across a wide area, mmWave base stations are primarily deployed to offer small-area coverage, serving a few users in each hotspot. Furthermore, the need for exceptionally high data throughput demands the utilization of wider transmission bandwidths. Supporting bandwidths of up to 400 MHz in a single carrier and over 1 GHz using carrier aggregation imposes even greater complexities on the implementation of RF and antenna elements.

9.4.4 Software-Defined Networking/SDN

Software-defined networking or SDN is a network model that separates the control plane from the data forwarding function (Nunes et al., 2014). This paradigm has been developed under the guidance of the non-profit operator-led consortium known as the Open Networking Foundation (ONF). In an SDN architecture, as illustrated in Fig. 9.7, network control is centralized within an SDN controller, while the underlying infrastructure is abstracted into a collection of forwarding elements. By centralizing network control, the SDN controller gains a comprehensive view of the entire network. This enables direct programmability of network control functions such as routing, congestion control, traffic engineering, and security inspection. The SDN controller acts as a centralized point for managing and configuring the network, facilitating greater flexibility and agility in network operations. SDN typically consists of three primary components that together form its structure:

- Application Layer: This layer serves as the highest level of abstraction within SDN. It encompasses applications and services that interact with the network. Examples include network management applications, traffic engineering tools, security applications, and network monitoring applications. These applications communicate with the SDN controller to define and manage network behavior.
- Control Layer: The control layer assumes responsibility for the centralized control and management of the network. It comprises an SDN controller, which is a software-based entity functioning as the core intelligence of the SDN architecture. The controller interacts with the applications and directly manages the underlying network devices (e.g., switches, routers) via a southbound interface (e.g., OpenFlow). It receives information on the network's state, makes decisions concerning network behavior based on the instructions or policies set by the applications, and accordingly programs the data plane.
- Data Layer: The data layer, also known as the forwarding plane, encompasses the physical or virtual network devices responsible for forwarding network traffic. These devices include switches, routers, and other networking equipment. In an SDN architecture, the control plane is separated from the data plane. This decoupling means that the control decisions made by the SDN controller are implemented by the data plane devices. The SDN controller programs the forwarding tables or rules on the switches or routers to determine how packets are forwarded through the network.

While SDN does not directly tackle the technical challenges associated with network control, it introduces fresh avenues for developing and implementing inventive solutions to address these issues. It achieves this by exposing the network as a service to SDN applications via the northbound interface. SDN applications provide instructions to the SDN controller, which then translates them into specific configuration commands for the underlying infrastructure using the OpenFlow protocol through the southbound interface. SDN brings forth numerous technical advantages, which include:

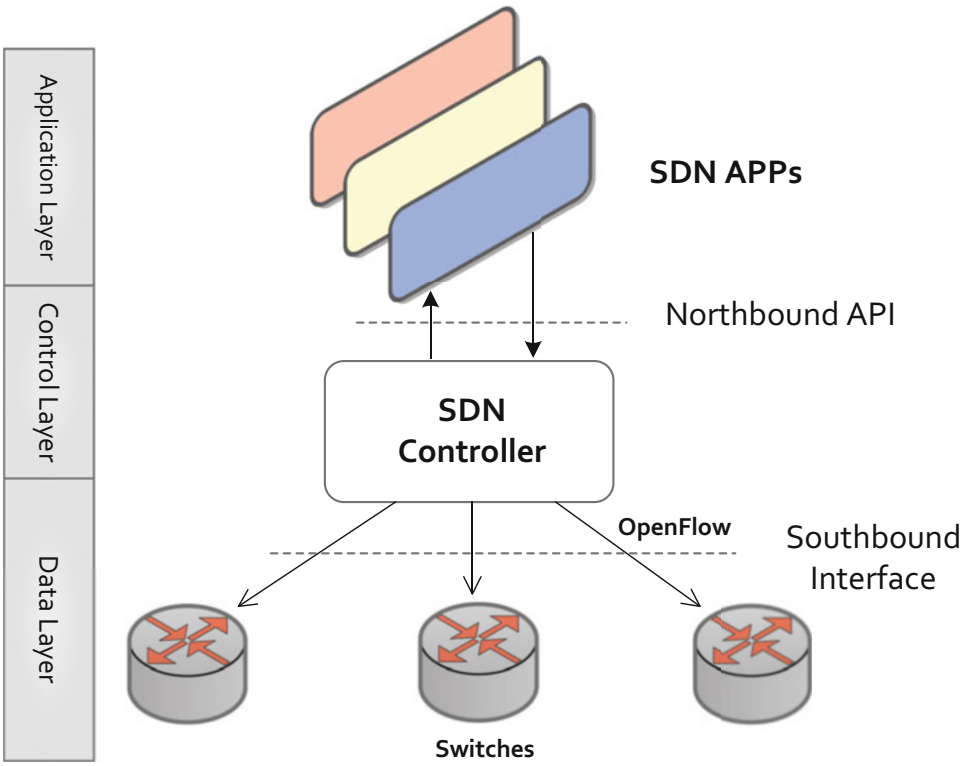


Fig. 9.7 The SDN architecture overview

- Centralized control of multi-vendor network elements
- Network automation and programmability by abstracting the underlying infrastructure and exposing it through a standard interface
- Rapid innovation through deploying new network applications and services without the need to configure specific elements or wait for vendor releases
- Increased network reliability and security due to centralized and automated management of network elements, uniform policy enforcement, and fewer configuration errors

9.4.5 Network Functions Virtualization/NFV

Network functions virtualization or NFV is a disruptive network paradigm that aims to revolutionize network deployment by separating software from hardware (Jiang et al., 2017b). It enables network operators to deploy network functions as virtualized software instances instead of relying on dedicated hardware appliances. These virtualized network functions can be executed on standard, general-purpose high-volume servers and can be dynamically migrated to different locations as needed, eliminating the need for new equipment installations. NFV offers a multitude of advantages, including:

- Low equipment costs and low power consumption through exploiting the economy of scale of the IT industry
- Faster time to market of new services
- Flexibility with elastic scale up and scale down of network capacity
- Multi-tenancy, which allows the sharing of a single platform for different applications, users, and tenants
- Enables border-independent software ecosystems and encourages openness

As illustrated in Fig. 9.8, the structure of NFV typically involves the following key components (Mijumbi et al., 2016):

- Network Functions: Network functions refer to the different tasks or functions performed by network appliances, such as routers, firewalls, load balancers, and intrusion detection systems. In NFV, these network functions are decoupled from dedicated hardware and implemented as software applications that can be executed on standard hardware servers.

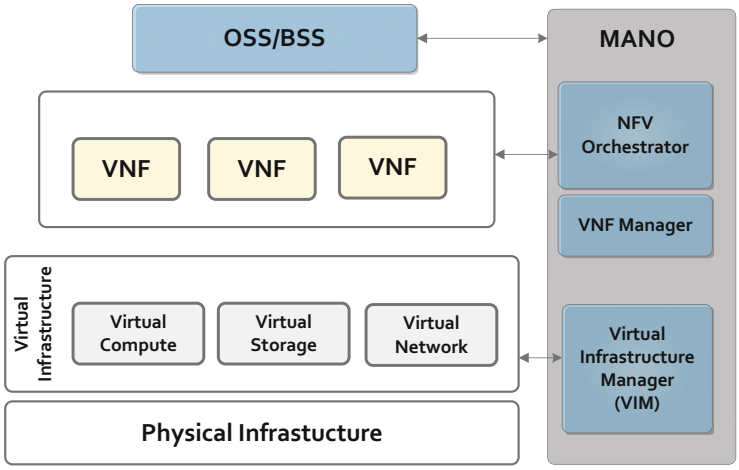


Fig. 9.8 The architecture of NFV-based systems

- NFV Infrastructure (NFVI): NFVI provides the underlying hardware and software resources necessary for hosting and running the virtualized network functions. It includes servers, storage devices, networking equipment, virtualization platforms (such as hypervisors), and management tools. NFVI provides the necessary compute, storage, and networking capabilities to support the virtualized network functions.
- Virtualized Network Functions (VNFs): VNFs are software-based instances of network functions that run on the NFVI. They represent the virtualized version of traditional network functions and are deployed and managed as software applications. VNFs can be instantiated, scaled, and managed dynamically based on the network's requirements, making them flexible and adaptable to changing network conditions.
- NFV Orchestrator: The NFV Orchestrator is responsible for managing and orchestrating the lifecycle of VNFs. It handles tasks such as VNF instantiation, scaling, termination, and coordination between different VNFs. The NFV Orchestrator interacts with the NFVI to allocate and manage the necessary resources for VNFs, ensuring optimal utilization and performance.
- Virtualized Infrastructure Manager (VIM): The VIM is responsible for managing the NFVI resources. It abstracts the underlying physical infrastructure and provides a virtualization layer to deploy and manage VNFs. The VIM interacts with the NFV Orchestrator to allocate resources, monitor the NFVI, and enforce resource policies.
- Management and Orchestration (MANO): MANO refers to the overall framework that combines the NFV Orchestrator, VIM, and other management components. It encompasses the management and orchestration of virtualized network functions, virtual infrastructure resources, and service lifecycle management. MANO ensures the efficient deployment, scaling, and operation of VNFs in the NFV environment.

Despite being developed independently by different organizations (ETSI and ONF), SDN and NFV are mutually beneficial and complement each other. On one hand, the SDN controller and SDN applications can be implemented as conventional VNFs and deployed on standard IT platforms. With the unified control provided by NFV MANO, software instances related to SDN can be dynamically instantiated, scaled, migrated, updated, and deployed within the virtualized infrastructure, offering flexibility and agility. On the other hand, NFV can leverage the network programmability facilitated by SDN to implement various network functions. By utilizing the capabilities of SDN, NFV can enhance its functionality and efficiency in delivering network services (Jiang et al., 2018). Together, SDN and NFV enable a synergistic approach that combines the benefits of software-defined networking and network functions virtualization, ultimately transforming network architectures and operations.

9.4.6 Network Slicing

Network slicing encompasses a range of technologies that enable the creation of specialized and dedicated logical networks within a shared physical network infrastructure. By employing tailored designs for functions, isolation mechanisms, and management tools, network slicing offers Network-as-a-Service (NaaS) capabilities, catering to diverse requirements from

various vertical industries. This allows for the provision of customized and self-contained virtual networks, known as network slices, based on the Service Level Agreement (SLA) between a mobile operator and a customer (Jiang et al., 2019). Each network slice represents an independent end-to-end logical network that operates with its own virtual resources, topology, traffic flow, and provisioning rules. However, these network slices utilize the same underlying physical infrastructure. Leveraging scalable resource allocation and flexible configuration options, a network slice can deliver personalized network capabilities such as data throughput, coverage, quality of service, latency, reliability, security, and availability.

There are various types of network slices to meet the specific communication needs of different users (Zhou et al., 2016). Major concepts for network slicing are as follows:

- *Network Slice Instance*: A network slice instance refers to a collection of network functions, whether shared or dedicated, and physical or virtual resources, all operating together to create a fully instantiated logical network. These network slices are designed to meet specific network requirements, such as ultra-reliability and low latency. Typically, a network slice instance spans multiple technical domains, including terminals, access networks, transport networks, core networks, and data centers that host third-party applications from various vertical industries.
- *Network Slice Type*: Network slice types serve as broad classifications for different network slice instances, representing specific demands for network solutions. In the context of 5G, three fundamental types have been recognized: eMBB, URLLC, and mMTC. These three types provide a foundational framework for network slicing in 5G, but it is possible for the list to expand in the future to accommodate evolving demands or advancements within the 5G ecosystem.
- *Tenant*: Tenants refer to customers of network slices, such as vertical industries or network operators, who utilize network slice instances to offer services to their own users. Consequently, tenants usually have distinct operation and management policies that are specifically tailored to their network slice instances. This allows tenants to maintain independent control over their respective network slice instances and ensure they align with their specific requirements and objectives.

9.4.7 Polar Codes

Over the past few decades, the primary challenge in the coding theory of digital communications has been to approach the limit of Shannon capacity while maintaining practical complexity. One significant advancement in this field is the introduction of coding randomness through the use of interleaving, which enables Turbo codes to achieve near-optimal performance with reasonable complexity. Turbo codes have found widespread application in 3G and 4G cellular systems such as WCDMA, CDMA2000, and LTE. Similarly, LDPC codes leverage coding randomness by incorporating pseudo-random connections between variable and check nodes. LDPC codes have demonstrated excellent performance, leading to their successful adoption in WiMAX specifications such as IEEE 802.16e and IEEE 802.16m. These coding techniques, Turbo codes and LDPC codes, have played a significant role in enabling reliable and efficient communication in various wireless systems.

In 2009, a breakthrough coding scheme called polar codes was introduced by Arikan, revolutionizing the construction of error-correcting codes and paving the way to achieve the Shannon capacity (Arikan, 2009). The key concept behind polar codes is channel polarization, which can be regarded as a Matthew effect in the digital world (*the rich gets richer and the poor gets poorer*). This phenomenon allows for the recursive transformation of multiple independent uses of a Binary-input Discrete Memoryless Channel (B-DMC) into a series of successive uses of synthesized binary-input channels. Initially, these independent channels undergo a transformation that results in two types of synthesized channels: good and bad channels. These channels are polarized, meaning they transmit a single bit with slightly varying reliability. This process forms the foundation of polar codes, capitalizing on the polarization phenomenon to achieve optimal transmission performance. Through the iterative application of polarization transformation on these channels, the mutual information of the synthesized channels tends to converge to two extremes: near-zero for the noisy channels and close to one for the noiseless channels. This property allows for the transmission of information bits over the noiseless channels while assigning frozen bits to the noisy ones, resulting in remarkable channel capacities.

In October 2016, Huawei made an announcement stating that they had achieved a downlink rate of 27 Gbps using polar codes. This breakthrough showcased that polar codes were capable of fulfilling all three usage scenarios defined by ITU IMT-2020: eMBB with transmission rates up to 20 Gbps, URLLC with 1 ms latency, and mMTC to support a massive number of connections. The successful implementation of polar codes demonstrated their effectiveness in providing efficient channel coding techniques for 5G networks, enabling significantly higher spectrum efficiency while maintaining a practical decoding complexity to minimize implementation costs. In November 2016, 3GPP approved the adoption of polar codes for the control

channel and LDPC codes for the data channel in 5G NR (Richardson & Kudekar, 2018). This decision further solidified the importance of polar codes as a key component in the 5G network architecture.

9.5 Summary

For previous generations, from 1G to 4G, the mission of cellular networks was to provide telecommunications services to subscribers. The most significant feature of Fifth Generation (5G), also known as IMT-2020, is centered around expanding the sphere of cellular communications, moving beyond connecting people to interconnecting humans, machines, and things. This chapter offers readers a comprehensive view of 5G, including its driving forces, three usage scenarios (enhanced mobile broadband, ultra-reliable low-latency communications, and massive machine-type communications), key performance indicators, and the standardization process. Through studying this chapter, readers gained the fundamentals of key technologies that empowered the implementation of 5G cellular systems. To offer a deep understanding, the next chapter will delve into the salient features and advancements introduced in the 3GPP Release 15, marking a pivotal transition from 4G to 5G technologies.

9.6 Exercises

1. Identify the big leap made by 5G in comparison with the previous generations of cellular systems. Hint: The first and second generations were designed for mobile voice services, while the following two generations were optimized for mobile data services.
2. In ITU-R M.2083 recommendations, three usage scenarios were defined. Which IMT-2020 usage scenarios do you know?
3. Describe the peak data rate specified for 5G.
4. What is the maximal bandwidth in 5G systems?
5. Which of the following technologies have been adopted by 5G?
   (A) Massive MIMO
   (B) Polar codes
   (C) Network slicing
   (D) Orthogonal frequency-division multiple access (OFDMA)
   (E) Single-carrier frequency-division multiple access (SC-FDMA)
6. What is the main difference between non-standalone (NSA) and standalone (SA) 5G?
7. Orthogonal multiple accesses, such as FDMA, TDMA, CDMA, and OFDMA, have been implemented in 1G to 4G systems, respectively. Each user is assigned to an orthogonal radio resource unit in terms of frequency, time, code, or a combination of these domains. In contrast, non-orthogonal multiple access/NOMA attracted a lot of attention during the era of 5G. What are the benefits brought by NOMA and what are its costs?
8. Identify the philosophy of software-defined networking or SDN.
9. Identify the philosophy of network functions virtualization or NFV.
10. In November 2016, 3GPP approved the adoption of polar codes as the channel coding technique for the control channel and LDPC codes for the data channel in 5G NR. Compare these two coding techniques.

New Radio (NR) and 5G Core Networks

10

The 3GPP Release 15, as mentioned in previous chapters, is the first 5G standard set that fully fulfills the requirements of IMT-2020. In this chapter we will focus on it to introduce the design and key features of the NG-RAN and 5GC, just like we introduced LTE-A regarding the 3GPP Release 10 in Chap. 8.

10.1 Deployment Options and Migration Paths

In 3GPP Release 15, various connectivity options have been delineated, falling into two broad categories: SA and NSA. SA deployment connects one independent RAN to either the EPC or the 5GC. Conversely, NSA deployment links both the E-UTRAN and the 5G NR access to the core network, a configuration known as multi-connectivity. Though the terminology such as #2, #3, etc. originated from the 5G study phase of 3GPP Release 15, it is not explicitly used in the 3GPP specifications. Options #1-#5 and #7 are illustrated in Fig. 10.1. Generally, Option #1 represents legacy LTE/LTE-A deployments; Option #3 integrates 5G NR as a secondary node into the legacy system for multi-connectivity, while Options #2, #4, #5, and #7 utilize the 5GC in various configurations.[1]

In all NSA options, the control plane is routed solely through the master node, with the user plane divided across the master and secondary nodes. Depending on the specific user plane traffic splitting, each of the NSA options #3/4/7 has more than one variant, as detailed in Fig. 10.2.

The choice between SA and NSA deployment options is pivotal for operators transitioning to 5G. Each option presents unique advantages and considerations, influenced by factors such as existing infrastructure, investment needs, targeted services, and strategic objectives. The comparative strengths and weaknesses of SA and NSA, as well as EPC and 5GC, are summarized in Table 10.1. Building on this analysis, different migration paths from 4G EPS to 5GS are explored in Table 10.2.

10.2 Network Slicing

Network slicing, a central feature of 5G architecture, facilitates the creation of multiple virtual networks atop a shared physical infrastructure. This section delves into the concept of network slicing, emphasizing its role as an E2E solution encompassing both the NG-RAN and the 5GC.

The 5G network slicing architecture is conceptualized through three interrelated layers: the Resource Layer, the Network Slice Instance (NSI) Layer, and the Service Instance Layer, as illustrated in Fig. 10.3. Each layer plays a distinct role in the overall architecture:

- **Resource Layer:** This foundational layer consists of physical and virtual resources, such as network infrastructure, radio resources, virtual resources, and network functions (NFs). It provides the underlying infrastructure upon which network slices are built. Resources can be dynamically allocated and managed to meet the specific needs of different network slices.

[1] Option #6, which had once been proposed to connect SA 5G NR to EPC, was later removed from the probable options and not specified in 3GPP Release 15.

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5_10

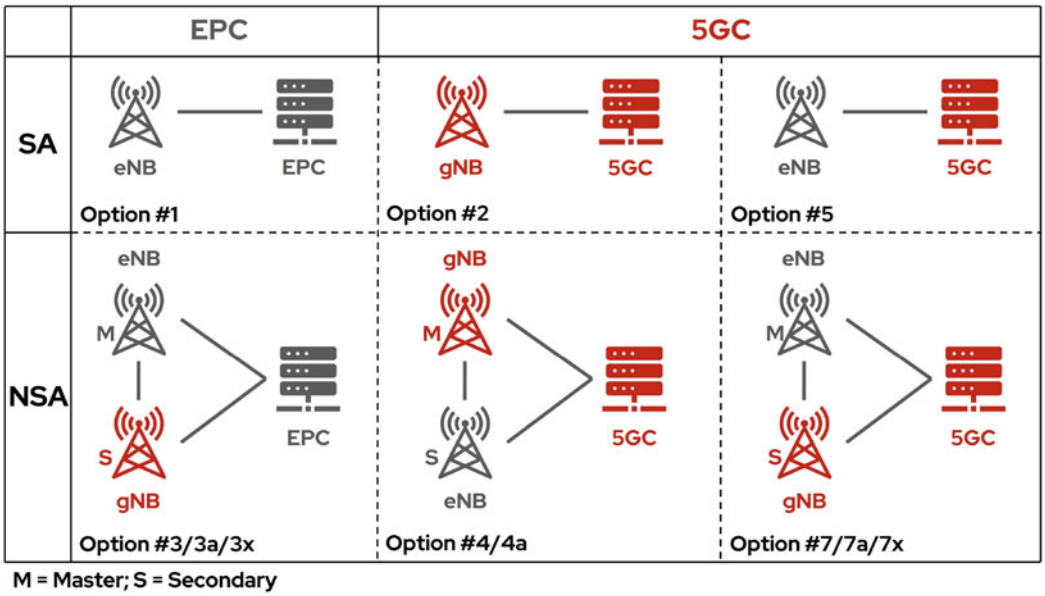


Fig. 10.1 5G deployment options

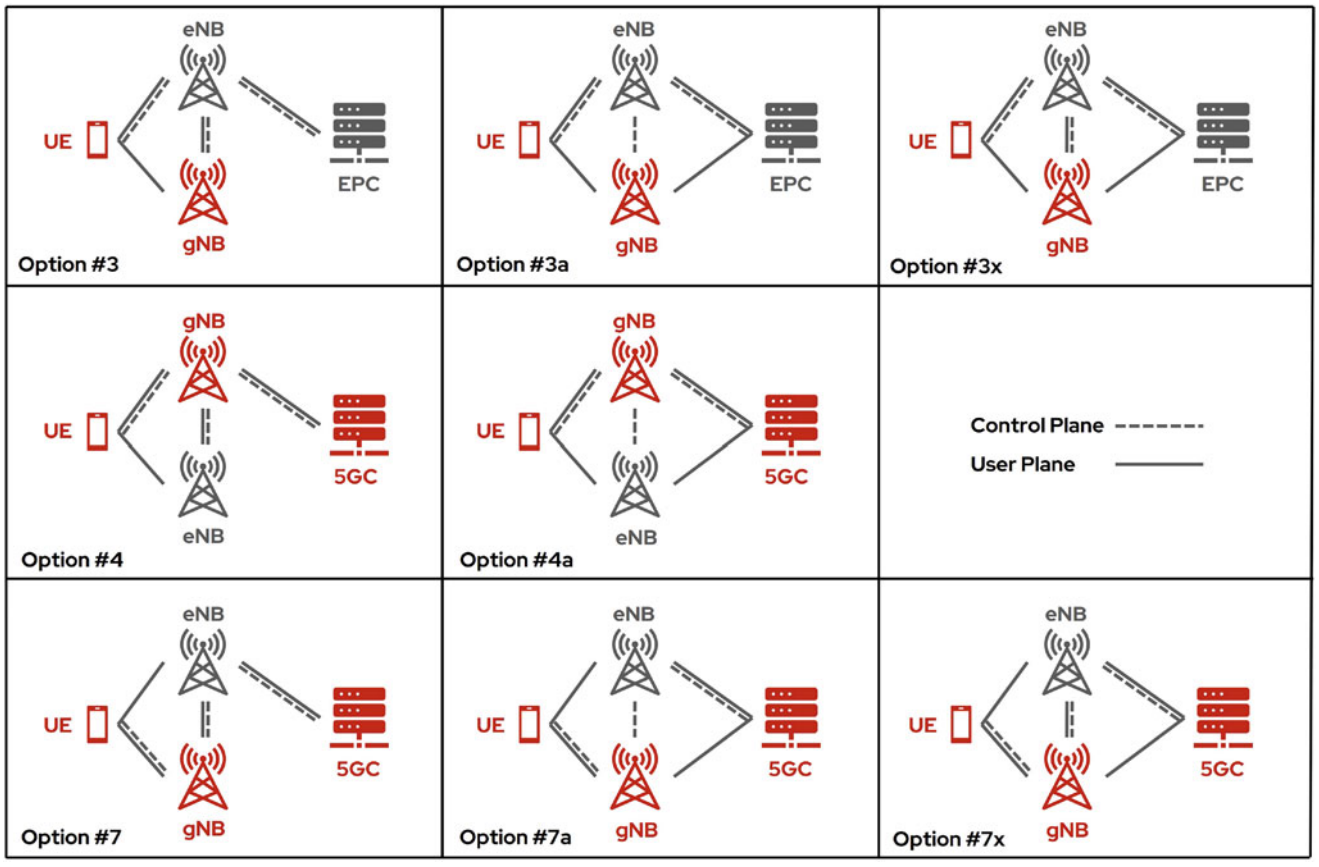


Fig. 10.2 Traffic flow in different NSA options

Table 10.1 Comparison of 5G radio access and core networks (Kim & Zarri, 2018)

		Advantages	Disadvantages
RAN	SA	• Simple to manage • Inter-generation handover between 4G and 5G	• Unable to leverage existing LTE deployments if NR is used in SA
	NSA	• Leverage existing LTE deployments	• Tight interworking between LTE and NR required • May impact end user experience
CN	EPC	• Leverage existing EPC deployment	• Cloud support is optional
	5GC	• Cloud native • Easier to support multiple access	• New deployment required

Table 10.2 Comparison of migration paths from 4G to 5G (Kim & Zarri, 2018)

Path	Use Case	Deployment	Device/Network	Voice
EPS to SA#2	+ Full 5G use cases	- 5GC benefits - Needs to retain EPC	+ Little impact on LTE	+ IMS voice supported
EPS to SA#3	- Limited support for 5G use cases	+ Leverage LTE + Quick time-to-market - No 5GC benefit	+ EPC procedures - Impact on LTE	+ Leverage existing VoLTE service
NSA #3 to NSA#7 / SA#5	+ Full 5G use cases - Initially limited	+ 5G Core benefits - Needs to retain EPC	- Impact on NR, LTE - Impact on IMS - 5GC deployment	+ IMS voice supported - No CS interworking from 5GS
NSA #3 to NSA#3 / SA#2	+ Full 5G use cases - Initially limited - Core migration	+ 5G Core benefits - Needs to retain EPC - Wide area NR	- Impact on NR, LTE - Impact on IMS - 5GC deployment	+ IMS voice supported - No CS interworking from 5GS
NSA #3 to NSA#4 / SA#2	+ Full 5G use cases - Initially limited - Core migration	+ 5G Core benefits - Needs to retain EPC	- Impact on NR, LTE - Impact on IMS - 5GC deployment	+ IMS voice supported - No CS interworking from 5GS

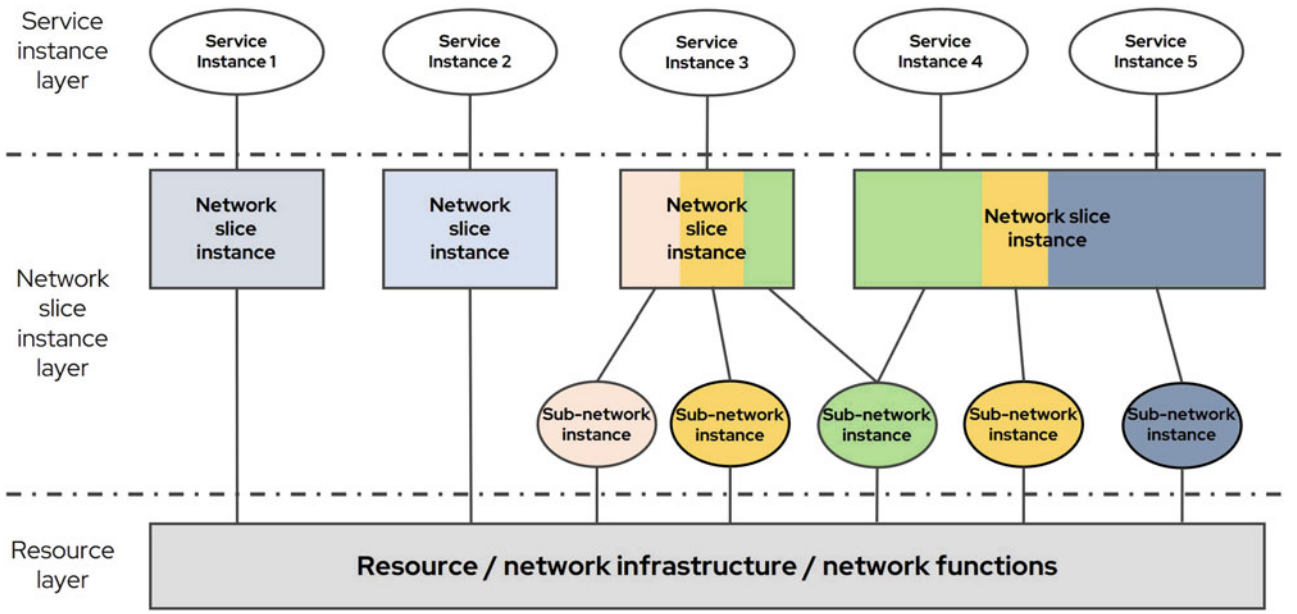


Fig. 10.3 Network slicing conceptual outline (Alliance, 2016a)

- **NSI Layer:** This layer represents complete logical networks, where NSI are created by the network operator using network slice blueprints. Each NSI can be independently tailored to match the requirements of a service instance. In addition, an optional entity called sub-network instance, which is a set of NFs, may be created based on a sub-network blueprint. A NSI may be composed of zero or more sub-network instances, which may be shared among multiple NSI.
- **Service Instance Layer:** This layer encapsulates the services and applications delivered to end users or business tenants, each associated with a specific NSI, ensuring tailored network resources.

It shall be noted that network slicing is an E2E concept. In 5G, it transcends both the NG-RAN and the 5GC, offering a seamless approach to network customization. In the RAN, slicing customizes radio resources, optimizing performance for specific services. Within the core network, slicing virtualizes and customizes core network functions, including dedicated instances of functions like the User Plane Function (UPF) or the Session Management Function (SMF). In both domains, dynamic creation, configuration, and optimization of network slices are vital for performance monitoring and real-time adjustments.

10.3 5G New Radio

10.3.1 Overview to the Key Features

The advent of 5G NR heralds a substantial evolution in capabilities and features, distinguishing it from its predecessor, E-UTRA. The key features include:

- **Extended Spectrum:** 5G NR extends into mmWave bands up to 52 GHz, including the 28 GHz and 39 GHz bands. This expansion enables higher data rates and capacity, optimizing spectrum utilization for bandwidth-intensive applications. The NR spectrum is divided into two frequency ranges (FRs): sub-6 GHz FR1 and over-6 GHz FR2, introducing several new frequency bands (see Table 10.3).
- **Increased Carrier Bandwidth:** 5G NR enhances the maximum carrier bandwidth to 100 MHz in sub-6 GHz bands and up to 1 GHz above 6 GHz, leading to higher peak data rates and improved spectral efficiency.
- **CA:** 5G NR supports aggregation of up to 16 component carriers, allowing overall transmission bandwidths up to 6.4 GHz. This flexibility in frequency band combination enhances network performance.
- **Flexible UE Channel Bandwidth:** 5G NR introduces flexibility in UE channel bandwidths within the same spectrum, supporting diverse deployment scenarios and non-contiguous spectrum allocations. This adaptability enhances spectrum efficiency, catering to various use cases with implications on RF requirements.
- **Advanced Beamforming:** 5G NR introduces dynamic analog beamforming and supports up to 12 layers of digital beamforming, improving signal quality and reducing interference for robust connections.
- **New Channel Coding:** 5G NR adopts LDPC for the user plane and Polar coding for the control plane, enhancing error correction, reliability, and efficiency.
- **Flexible Subcarrier Spacing:** 5G NR supports multiple subcarrier spacings, enhancing spectral efficiency and accommodating a wide array of use cases.
- **Self-Contained Subframe:** 5G NR can implement self-contained subframes, reducing latency and enhancing real-time application support.
- **Enhanced Spectrum Occupancy:** 5G NR utilizes up to 98% of the channel bandwidth, maximizing spectral efficiency for higher data rates and better spectrum utilization.

These collective advancements contribute to 5G NR's superior performance, flexibility, and efficiency, unlocking a wide array of new use cases and services.

Table 10.3 5G NR new bands, "o" for configurable, "-" infeasible, and "*" optional.

Frequency range	Frequency band		Subcarrier spacing (kHz)	Component carrier bandwidth (MHz)														
	Band number	Frequency band (GHz)		5	10	15	20	25	30	40	50	60	70	80	90*	100	200	400*
FR1	n77/n78	3.7	15		o	o	o			o	o						–	–
			30		o	o	o			o	o	o		o	o	o	–	–
			60		o	o	o			o	o	o		o	o	o	–	–
	n79	4.5	15							o	o						–	–
			30							o	o	o		o		o	–	–
			60							o	o	o		o		o	–	–
FR2	n257	28	60	–	–	–	–	–	–	–	o	–	–	–	–	o	o	
			120	–	–	–	–	–	–	–	o	–	–	–	–	o	o	o

Table 10.4 Features of 5G NR numerologies 0–4

Numerology	0	1	2	3	4
Symbols per slot	14	14	14 (normal CP) 12 (extended CP)	14	14
Slots per frame	10	20	40	80	160
Slots per subframe	1	2	4	8	16
Subcarrier spacing	15 kHz	30 kHz	60 kHz	120 kHz	240 kHz
Symbol duration	66.7 µs	33.3 µs	16.7 µs	8.33 µs	4.17 µs
CP duration	4.7 µs	2.3 µs	1.2 µs (normal CP) 4.13 µs (extended CP)	0.59 µs	0.29 µs
TTI (slot length)	1 ms	500 µs	250 µs	125 µs	62.5 µs
Max. nominal system bandwidth	50 MHz	100 MHz	100 MHz (sub-6 GHz) 200 MHz (mmWave)	400 MHz	400 MHz
Max. FFT size	4096	4096	4096	4096	4096
Supported for data	Yes	Yes	Yes	Yes	No
Supported for sync	Yes	Yes	No	Yes	Yes
PRACH	Short preamble	Short preamble	Short preamble	Short preamble	N/A

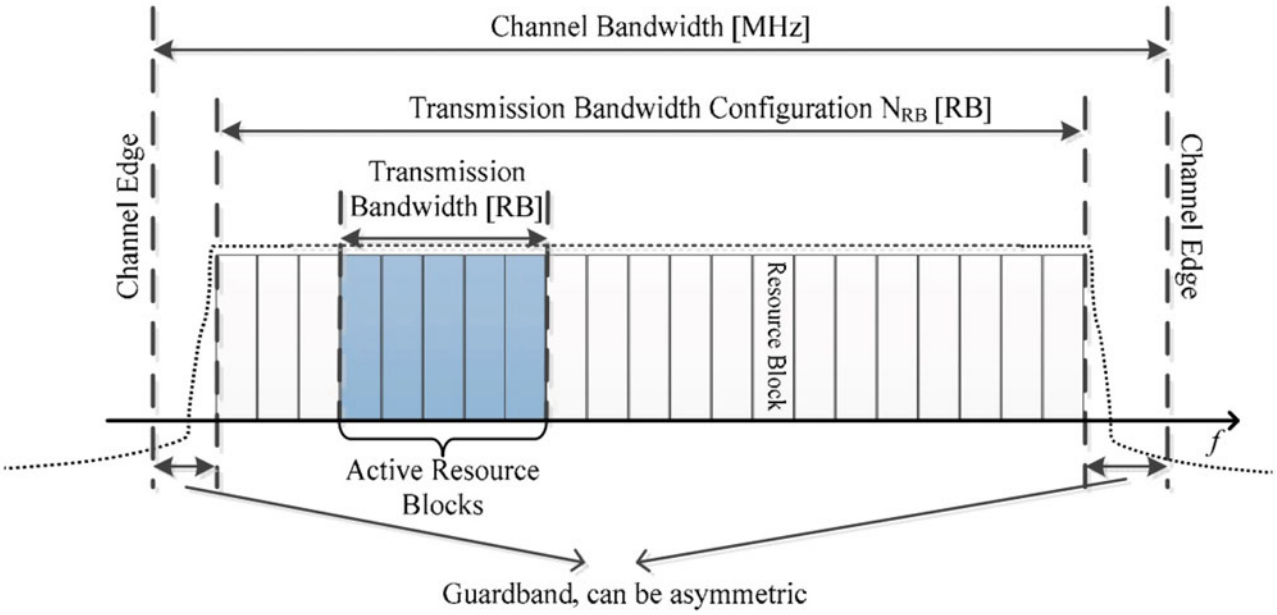


Fig. 10.4 Definition of the channel bandwidth and the maximum transmission bandwidth configuration for one NR channel (3GPP TS 38.101-1, 2023)

10.3.2 Waveform and Radio Frame Design

The evolution from E-UTRA to 5G NR has brought about significant advancements in waveform design. These enhancements not only improve the spectral efficiency and flexibility of the system but also enable coexistence with various technologies and adapt to diverse deployment scenarios.

As aforementioned, 5G NR introduces support for flexible subcarrier spacing RANging from 15 kHz to 240 kHz, which defines multiple possible numerologies, as listed in Table 10.4.[2] The introduction of variant numerologies allows using flexible TTI, which RANges between 62.5 µs and 1 ms, to support various use cases. More specifically, shorter TTI down to 62.5 µs plays a key role in latency-intolerant use cases such as URLLC. In contrast, E-UTRA's fixed subcarrier spacing of 15 kHz and TTI of 1 ms limited its ability to adapt to various use cases and deployment needs. The NR concepts of channel bandwidth and maximum transmission bandwidth configuration N_{RB} are illustrated in Fig. 10.4, and N_{RB} is specified as Table 10.5 shows.

[2] Only listing out numerologies 0–4, since 5 and 6 were later introduced in Release 17.

Table 10.5 The maximum transmission bandwidth configuration N_{RB} for each UE channel bandwidth and subcarrier spacing

Subcarrier spacing	UE channel bandwidth (MHz)											
(kHz)	5	10	15	20	25	30	40	50	60	80	90	100
15	25	52	79	106	133	160	216	270	N/A	N/A	N/A	N/A
30	11	24	38	51	65	78	106	133	162	217	245	273
60	N/A	11	18	24	31	38	51	65	79	107	121	135

Table 10.6 Usage of channel coding scheme in 5G NR

	TrCH/control information	Coding scheme
UP	UL-SCH	LDPC
	DL-SCH	
	PCH	
	BCH	Polar code
CP	DCI	Polar code
	UCI	Block code
		Polar code

Another significant change in 5G NR is the consistent use of cyclic prefix OFDM (CP-OFDM) across DL and UL, while DFT-s-OFDM is still available for UL as a complementary option. This contrasts with E-UTRA, where CP-OFDM is used in DL, and DFT-s-OFDM is used in UL. The use of CP-OFDM in 5G NR's UL overcomes the single-stream limitation of DFT-s-OFDM in E-UTRA, enabling MIMO support in UL. The PAPR issue, which motivated the use of DFT-s-OFDM in E-UTRA's UL, is mitigated in 5G NR through advanced power control techniques.

In addition, the flexibility in guard band and roll-off configuration in 5G NR also represents a significant advancement over E-UTRA. While E-UTRA had fixed guard bands and roll-off factors, 5G NR allows dynamic adjustments based on deployment scenarios, frequency bands, and numerology. This contributes to more efficient spectrum utilization, improved coexistence with other technologies, and adaptation to diverse scenarios. More specifically, the minimum width of guardband in 5G NR can be calculated as

$$\mathrm{GB}_{\mathrm{channel}} = (\mathrm{BW}_{\mathrm{channel}} \times -N_{\mathrm{RB}} \times \mathrm{SCS} \times 12)/2 - \mathrm{SCS}/2, \tag{10.1}$$

where guard bandwidth $\mathrm{GB}_{\mathrm{channel}}$, the channel bandwidth $\mathrm{BW}_{\mathrm{channel}}$, and the subcarrier spacing SCS are all measured in kHz.

10.3.3 PHY Layer Enhancements

Modulation and Coding Schemes Similar to E-UTRA, 5G NR supports up to 256-QAM in both UL and DL. To achieve better spectral efficiency, as referred in Sect. 10.3.1, 5G NR deploys new channel coding schemes, as briefly listed in Table 10.6. Detailed specifications of the coding schemes are provided in (3GPP TS 38.212, 2022).

Flexible Duplex There is a crucial difference between 5G NR and E-UTRA regarding the duplex. LTE/LTE-A, as we have introduced in Chap. 8, has two different radio frame structures for FDD and TDD, respectively. The PHY layer design and radio interface procedures were therefore also specified separately for the FDD and TDD versions of LTE. In contrast, 5G NR has a uniform frame structure that supports both FDD and TDD. Moreover, when operating in TDD, the time division between UL and DL is fixed in LTE/LTE-A and cannot be adjusted over time. In contrast, 5G NR supports dynamic TDD, where the time domain resources can be flexibly assigned and dynamically reassigned between UL and DL.

Massive MIMO and Beamforming LTE/LTE-A supports up to 8×8 MIMO and applies beamforming only on UP radio bearers. In 5G NR, massive MIMO up to the dimension of 256×64 is supported, while beamforming is applicable also for synchronization signals and control channels. Moreover, associated with mmWave technologies, 5G NR introduced analog beamforming into the RF domain to provide broad control of the beams, with digital beamforming working in conjunction to provide fine-tuning of the beams. In comparison, LTE/LTE-A only uses digital beamforming.

Table 10.7 Reference signals in 5G NR

Reference signal	Transmitted over	Functionality
UE-specific DM-RS	PDCCH	Required for demodulation
	PUCCH	
	PDCCH	
	PUSCH	
CSI-RS	[1]	CSI acquisition, beam management
SRS	[2]	
PBCH DM-RS	PBCH	Required for demodulation
PT-RS	PDCCH	Used for phase tracking
	PUSCH	
TRS	[3]	Used for time tracking

[1] The CSI-RS are transmitted only in DL, over their own dedicated resources, separate from the data or control channels
[2] The SRS are transmitted only in UL, over their own dedicated resources, separate from the data or control channels
[3] The TRS does not exist independently. A specific CSI-RS configuration is used as TRS

Reference Signals A significant renovation in the reference signal design is made in 5G NR. Unlike LTE/LTE-A, where the CRS is always transmitted for channel estimation regardless of the traffic, NR relies on user specific DM-RS to estimate the channels. For each UE, no DM-RS will be not transmitted unless there is data to transmit, so that the energy efficiency is improved. Moreover, to enable accurate beamforming and to support channel tracking, NR introduces various new types of RS, which are transmitted in both DL and UL. In addition to the DM-RSs, other types of UE-specific RS are including the CSI-RS, the Sounding Reference Signal (SRS), the Phase Tracking Reference Signal (PT-RS), and the Tracking Reference Signal (TRS). In addition, due to the removal of CRS, a Physical Broadcast Channel (PBCH) DM-RS is introduced to support PBCH decoding. More details are listed in Table 10.7.

10.3.4 Layer 2/Layer 3 Enhancements

The evolution from E-UTRAN to 5G NR has brought about significant enhancements in the L2 and L3 protocols, reflecting the system's increased complexity and the need for more flexible and efficient mechanisms.

The Service Data Adaption Protocol (SDAP) Sublayer The 5G NR radio interface protocol stack introduces a new L2 sublayer called SDAP. As depicted in Figs. 10.5 and 10.6, this sublayer lays only in the UP over the PDCP sublayer and maps QoS flows to data radio bearers. It ensures that the QoS requirements of different applications are met, providing a more granular control over the treatment of user data.

The introduction of the SDAP sublayer has several advantages. First, by dedicating the SDAP sublayer solely to QoS management, 5G NR can ensure that QoS requirements for different applications are met with higher precision. This is particularly crucial in 5G, which promises to support a myriad of applications with varying QoS needs, from eMBB to mMTC and URLLC. With the SDAP sublayer handling the QoS profiles for each flow, the system can make more informed decisions about resource allocation. Second, by offloading QoS management from the PDCP to the SDAP, the PDCP sublayer is simplified so as to focus on its primary responsibilities of header compression and security. This clear separation of duties between protocol layers or sublayers leads to better maintainability, easier troubleshooting, and a reduction in potential complexities. Third, a standalone SDAP sublayer provides a flexible and scalable architecture that can adapt to varying QoS needs of different services. With 5G's emphasis on network slicing, the role SDAP role in managing QoS profiles becomes even more critical, ensuring that each slice meets its specific QoS requirements.

Introduction of the New RRC State RRC_INACTIVE 5G NR introduces a new RRC state called RRC_INACTIVE, adding to the existing RRC_IDLE and RRC_CONNECTED states. This new state aims to provide a balance between energy efficiency and connection responsiveness.[3]

[3] The legacy E-UTRA system has no such state; however, an E-UTRA RRC_INACTIVE state is introduced in Release 16, as shown in Fig. 10.7.

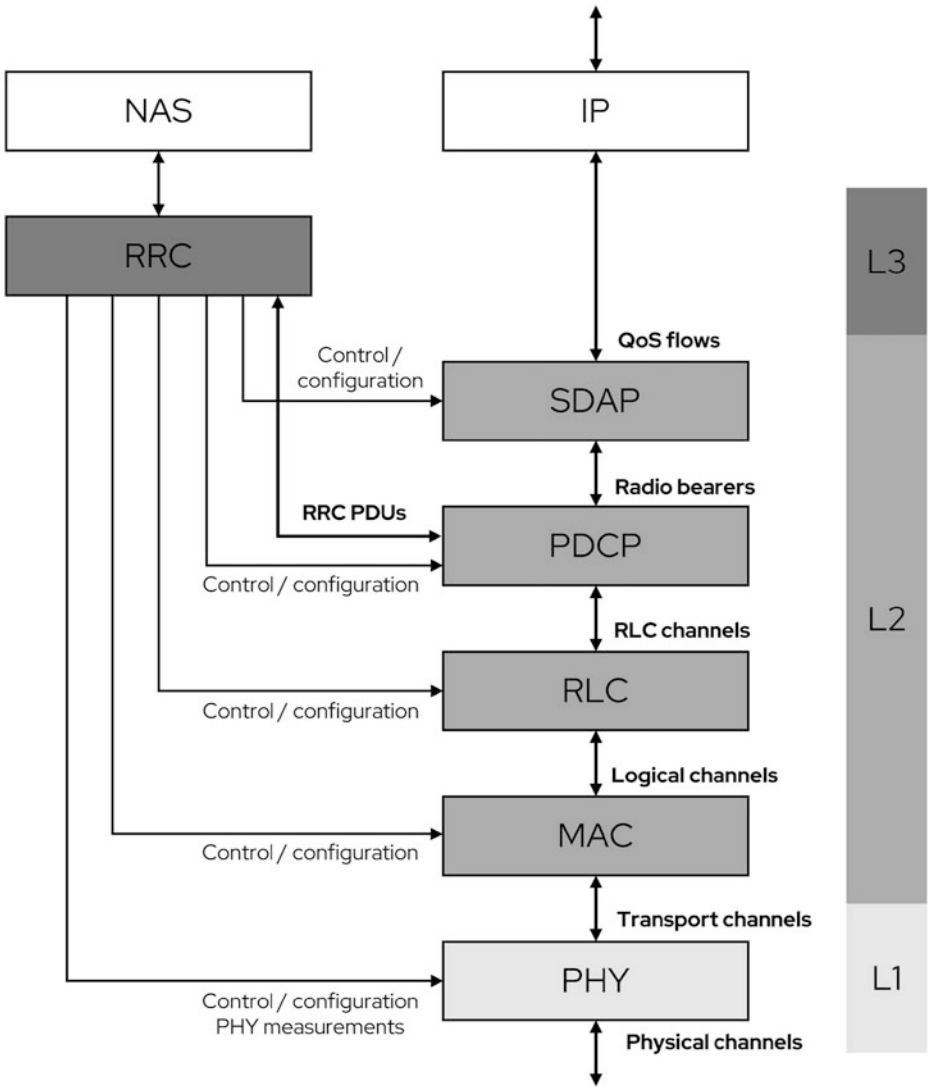


Fig. 10.5 5G NR radio interface protocol stack

In the RRC_INACTIVE state, the UE maintains its RRC context and some NG-RAN context but releases the radio resources. This allows for quicker reactivation of the connection compared to the RRC_IDLE state, where all contexts are released. The RRC_INACTIVE state thus enables faster transition to the RRC_CONNECTED state, reducing the connection setup time and enhancing the user experience for applications that require frequent but intermittent data transmission.

Enhancements in HARQ Mechanism 5G NR also brings about significant enhancements in the HARQ mechanism compared to E-UTRAN. The HARQ in 5G NR supports up to 16 parallel processes, allowing more flexibility in handling retransmissions and improving the system's robustness.

Furthermore, 5G NR introduces asynchronous HARQ for UL, enabling more efficient utilization of resources and better alignment with the DL and UL scheduling. This asynchronous operation allows for more flexible timing of feedback and retransmission, adapting to varying network conditions and user requirements.

The HARQ enhancements in 5G NR are integral to achieving the system's goals of higher reliability, lower latency, and more efficient use of resources. The L2 functions, including HARQ, are illustrated in Fig. 10.8.

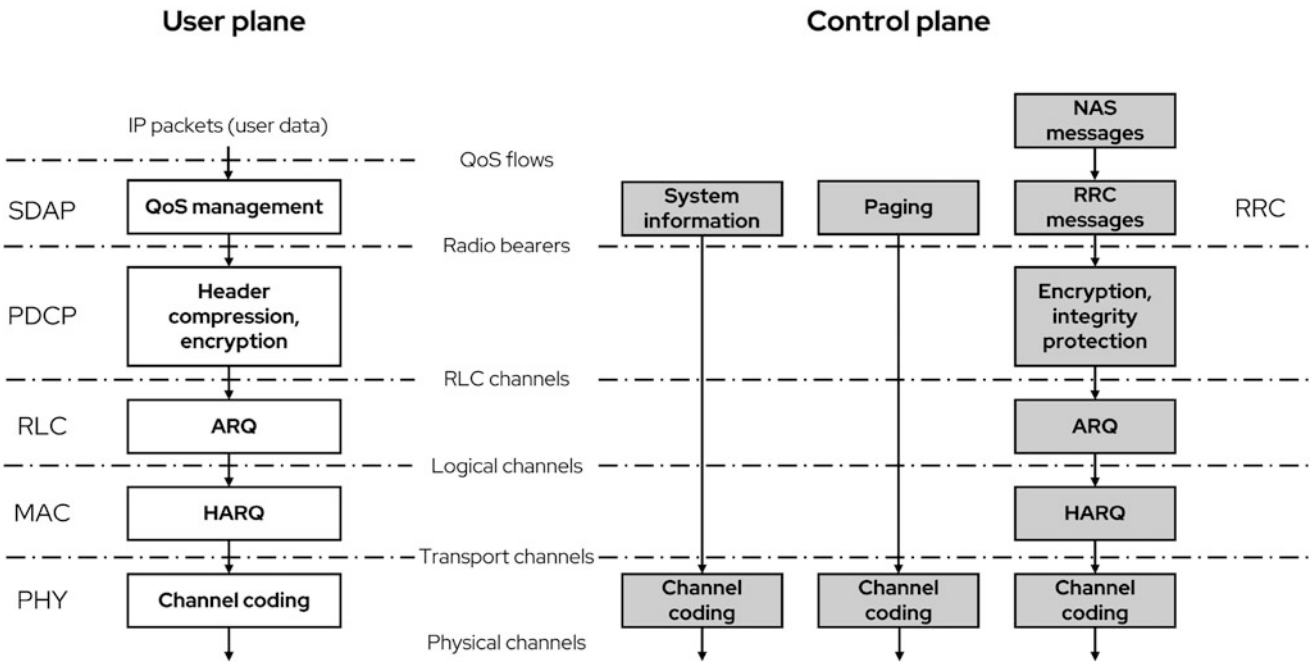


Fig. 10.6 5G NR radio interface functions in UP and CP

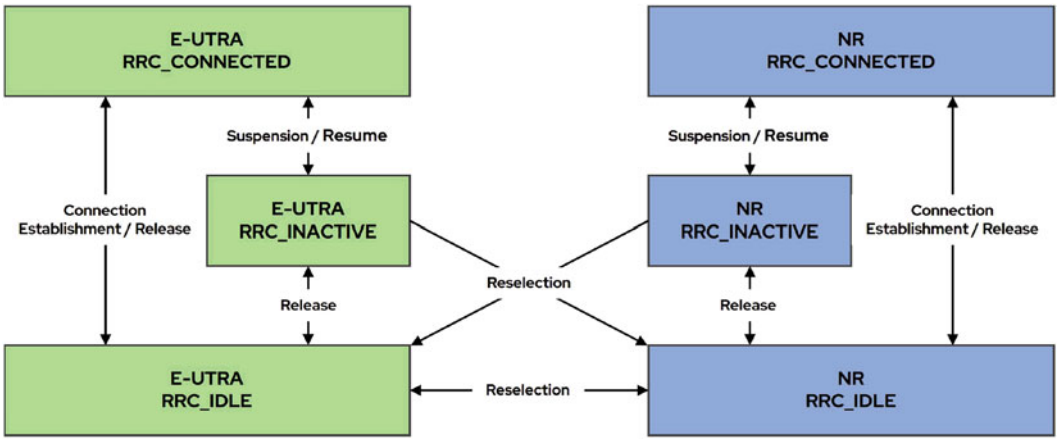


Fig. 10.7 RRC status and transitions in NR coexisting with E-UTRAN

10.3.5 Coexistence with Legacy RATs

5G NR is designed to coexist seamlessly with legacy technologies, ensuring a smooth transition and maximizing the utilization of existing infrastructures.

E-UTRA-NR Dual Connectivity (EN-DC)
In 5G NR, EN-DC is a key feature that enables seamless coexistence with LTE/LTE-A networks. EN-DC allows for the configuration of a Master Cell Group (MCG) and a Secondary Cell Group (SCG). The MCG is typically E-UTRA, and the SCG is NR. This dual connectivity ensures that UE can maintain high data rates and reliability by leveraging both E-UTRA and NR resources.

Dynamic spectrum sharing (DSS) DSS is another feature specified in 5G NR to facilitate coexistence with legacy RATs. It allows 5G NR and E-UTRA to share the same frequency band dynamically. The standard outlines the conditions under which the spectrum can be shared, such as quality of service requirements and network load conditions.

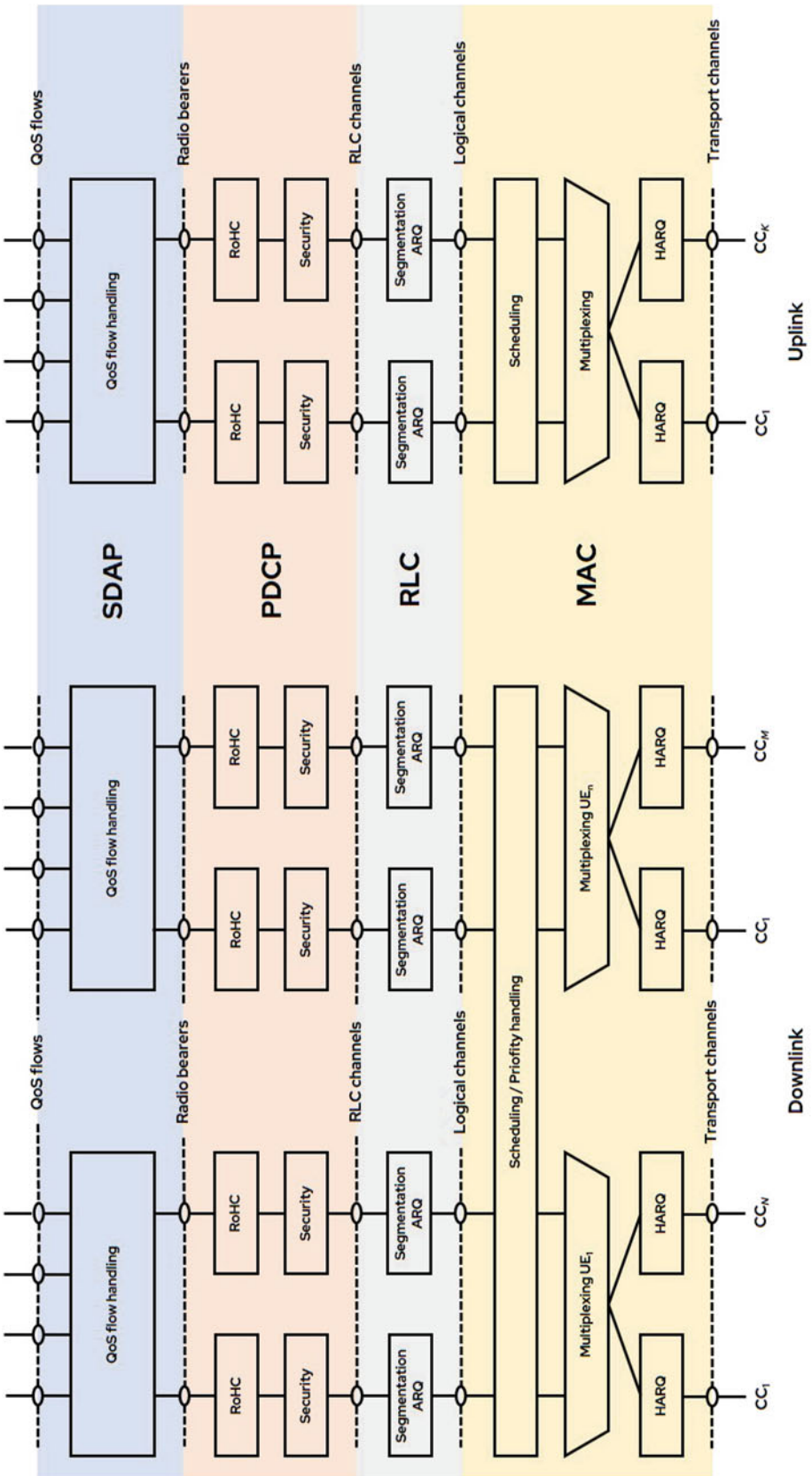


Fig. 10.8 5G NR L2 functions in DL and UL, RoHC = robust header compression

Non-Licensed Bands 5G NR also introduces the ability to operate in non-licensed bands, commonly referred to as NR Unlicensed (NR-U). This allows 5G NR to coexist with Wi-Fi and other technologies that operate in the same spectrum. Specific channel configurations and Listen-Before-Talk (LBT) mechanisms are defined to ensure fair coexistence in shared spectrum environments.

10.4 NG-RAN Architecture

The NG-RAN serves as the radio access backbone of the 5G system, engineered to meet the diverse and evolving requirements of next-generation wireless networks. Its architecture is inherently flexible and scalable, offering specialized configurations like Centralized RAN (C-RAN) for enhanced efficiency. Advanced features such as RAN slicing further enable a customizable network infrastructure.

10.4.1 Overall Architecture and Interfaces

Inheriting the flat structure of its predecessor, E-UTRAN, NG-RAN is streamlined with Next-Generation Node B, or gNB (gNodeB) as its solitary entity, akin to the eNodeB in E-UTRAN. This design choice simplifies the architecture, reducing the number of hierarchical layers and thereby enhancing efficiency. However, a pivotal innovation sets NG-RAN apart: the gNodeB is architecturally divided into the Centralized Unit (CU) and the Distributed Unit (DU), as illustrated in Fig. 10.9. This functional split offers a multitude of advantages, including but not limited to, scalability, lower latency, and more efficient resource allocation.

The CU is responsible for control plane functionalities, while the DU handles the user plane, allowing for a more flexible and optimized deployment. This separation is particularly advantageous for operators, as it enables them to tailor their networks to specific use cases or operational requirements.

In terms of interfaces, NG-RAN introduces three key elements: NG, Xn-C, and F1. The NG interface connects the gNodeB to the 5GC, facilitating seamless communication between the two. The Xn-C interface is employed for control plane signaling between different gNodeB, ensuring coordinated operations. Lastly, the F1 interface is designed for communication between the CU and DU, allowing for a harmonized operation of the split architecture.

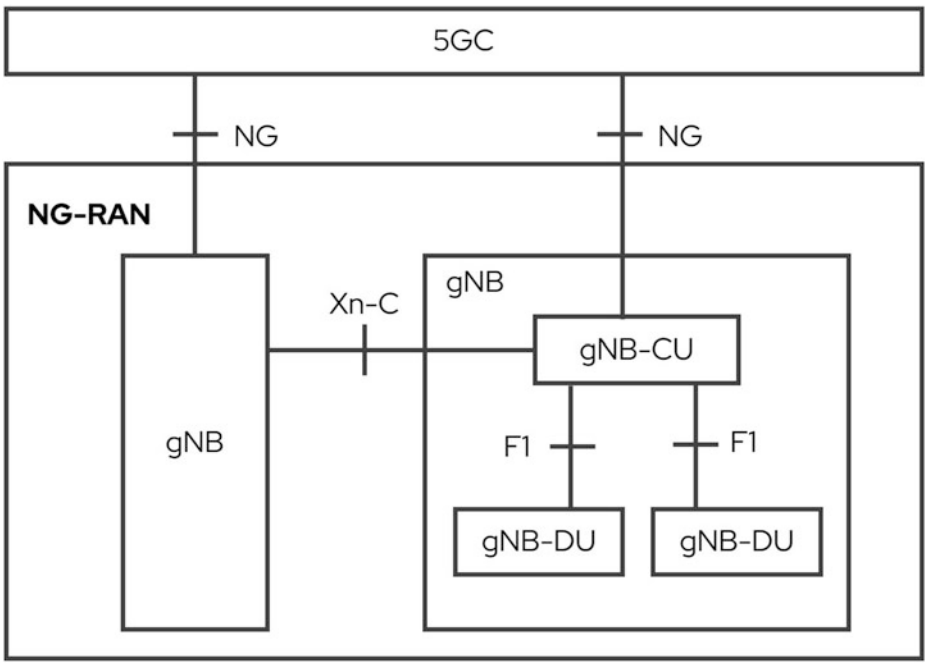


Fig. 10.9 Overall NG-RAN architecture (3GPP TS 38.401, 2020)

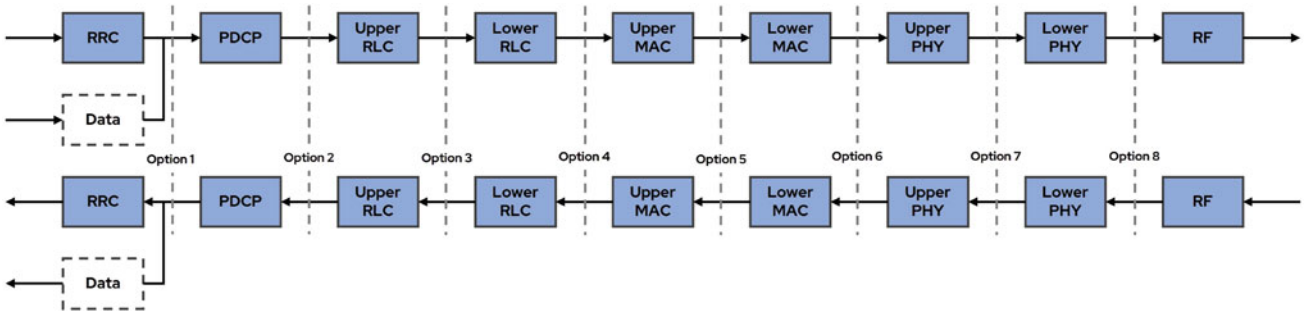


Fig. 10.10 Functional split options between CU and DU (3GPP TR 38.801, 2017)

This architectural evolution not only meets the diverse requirements of 5G but also lays the groundwork for future advancements in wireless communication networks.

10.4.2 Functional Splitting

The architectural division of the CU and DU in NG-RAN is a critical ingredient for achieving unprecedented deployment flexibility. Unlike legacy networks, which often had rigid architectural constraints, NG-RAN allows for a RANge of functional splitting options between the CU and DU. This adaptability is crucial for operators to tailor their networks to specific operational requirements, use cases, or even future technologies.

The choice of functional split impacts several key performance indicators, such as latency, resource efficiency, and system complexity. Generally, a split closer to the CU side results in higher latency but allows for more efficient resource pooling and reduced complexity. Conversely, a split closer to the DU side offers lower latency at the expense of resource efficiency and increased complexity.

Options 1 through 8 offer varying degrees of functional splits, as illustrated in Fig. 10.10 and compared in Table 10.8. These options RANge from a 1A-like split, where only the RRC resides in the CU, to a PHY-RF split, where only the RF functionality is in the DU. Each option has its own merits and drawbacks, allowing operators to select the most suitable configuration for their specific needs.

Special suboptions, such as 3-1, 3-2, 7-1, 7-2, and 7-3, offer even more granular control over the functional split. For instance, Option 3-1 and 3-2 allow for different compositions of the RLC, while Options 7-1, 7-2, and 7-3 provide various configurations for the physical layer functions. These suboptions are particularly useful for specialized deployments that require fine-tuned performance characteristics.

It is worth noting that Option 4 is rarely considered in practical deployments, as it offers no discernible benefits over the other options.

10.4.3 RAN Architectural Evolution: C-RAN, V-RAN, and O-RAN

As we delve deeper into the intricacies of NG-RAN, it becomes imperative to explore its alignment and potential evolution with other transformative RAN architectures like C-RAN, Virtualized RAN (V-RAN), and Open RAN (O-RAN). These architectures not only encapsulate unique advantages but also signify the evolutionary paths that NG-RAN could traverse.

Alignment with C-RAN Originating from the need to optimize resource utilization and reduce operational costs, C-RAN emerged as a paradigm shift in network architecture. It is characterized by centralized Baseband Processing Units (BBUs) and distributed Remote Radio Units (RRUs). This centralization allows for more efficient resource pooling and facilitates advanced features like CoMP transmission. The NG-RAN architecture, with its distinct CU and DU components, naturally aligns with these C-RAN principles, offering similar deployment flexibility and advanced feature sets.

Table 10.8 Summary on characteristics of different CU–DU split options (3GPP TR 38.801, 2017)

	Opt. 1	Opt. 2	Opt. 3-2	Opt. 3-1	Opt. 5	Opt. 6	Opt. 7-3	Opt. 7-2	Opt. 7-1	Opt. 8
Baseline available	No	Yes (LTE DC)	No							Yes (CPRI)
Traffic aggregation	No	Yes								
ARQ location	DU			CU (May be robuster under non-ideal transport conditions)						
CU Resource pooling	Lowest	In between (higher on the right)								Highest
	RRC	RRC + L2 (partial)				RRC + L2	RRC + L2 + PHY (partial)			RRC + L2 + PHY
Transport NW latency req.	Loose				Note 4	Tight				
Transport NW peak BW req.	N/A	Lowest	In between (higher on the right)							Highest
	No UP req.	Baseband bits						Quant. IQ (f)		Quant. IQ (t)
	–	Scales with MIMO layers							Scales with antenna ports	
Multi-cell/freq. coordination	multiple schedulers (independent per DU)				Centralized scheduler (can be common per CU)					
UL adv. Rx	Note 4						N/A	Note 4	Yes	
Remarks	Note 1				Note 2/3	Note 2				

Note 1: May be beneficial for URLLC/MEC
Note 2: Complexity due to separation of scheduler & PHY processing
Note 3: Complexity due to separation of scheduler & HARQ
Note 4: Was not clarified during the study phase of 3GPP TR 38.801 (2017)

Potential for V-RAN Integration V-RAN was motivated by the need for greater network agility and scalability. By decoupling network functions from dedicated hardware and running them as software on commercial off-the-shelf hardware, V-RAN offers unprecedented flexibility and dynamic resource allocation. These principles are in harmony with the inherent adaptability of NG-RAN, making it a prime candidate for V-RAN integration in future network evolutions.

Evolution Toward O-RAN The O-RAN initiative was born out of the desire for more open and interoperable network ecosystems. Driven by the O-RAN Alliance, this architecture aims to break vendor lock-in and foster innovation by promoting modular RAN systems with components from different vendors. While NG-RAN is the primary focus of 3GPP Release 15, its architecture's inherent flexibility and modular design make it well-suited for potential integration or compatibility with O-RAN in future 3GPP releases.

By examining the origins and motivations behind these architectures, we can better appreciate how NG-RAN serves as both an endpoint and a stepping stone toward more flexible, open, and efficient RAN architectures.

10.4.4 RAN Slicing in NG-RAN

Network slicing is a primary aspect of 5G and beyond, enabling the creation of multiple logical networks on top of a shared physical infrastructure. While network slicing is an E2E concept, RAN slicing in NG-RAN under 3GPP Release 15 introduces unique features tailored specifically for the radio access network.

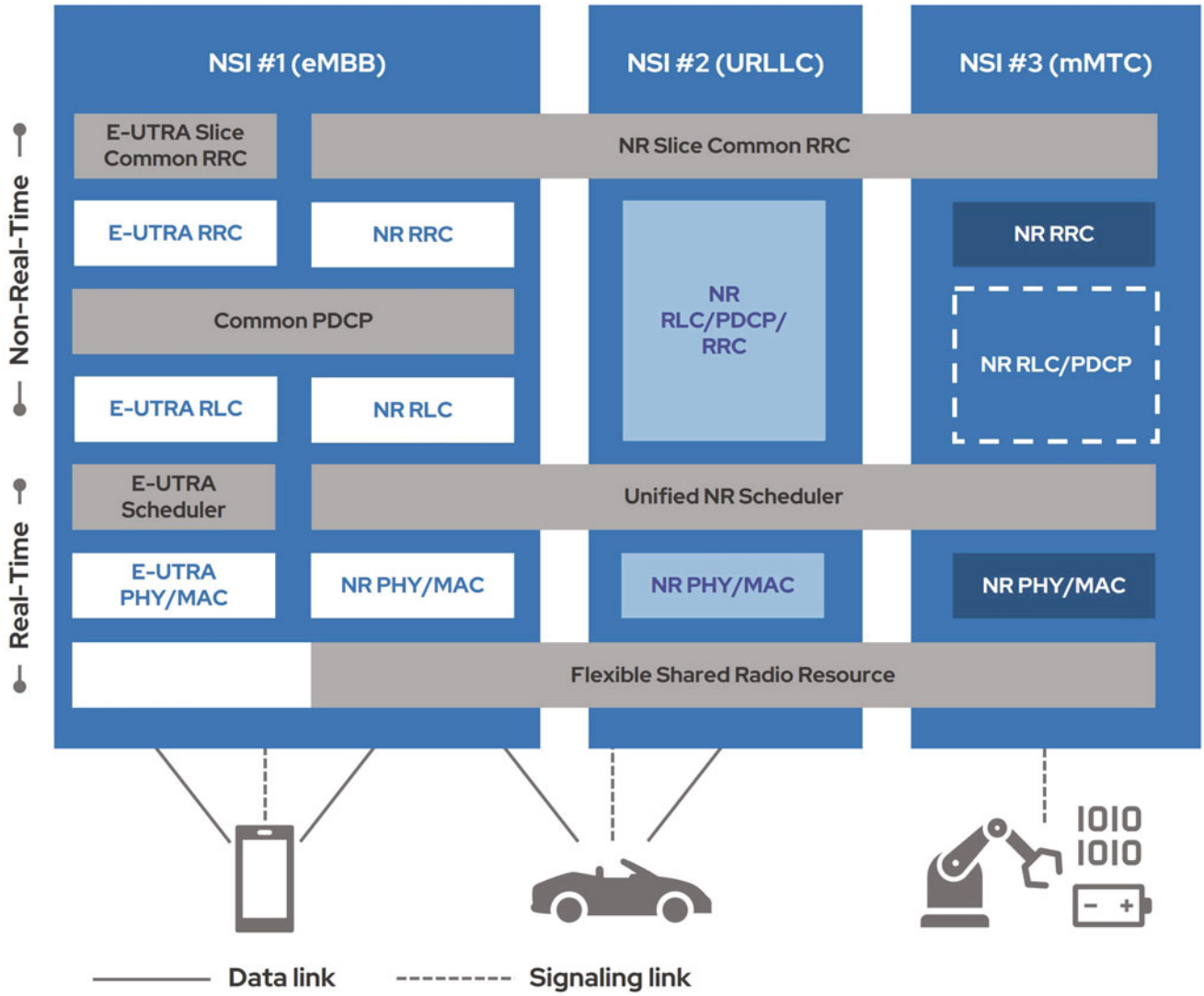


Fig. 10.11 RAN architecture with network slicing support example (China Mobile Communications Corporation et al., 2017)

Tailoring NSI for Different Use Cases RAN slicing allows for the creation of NSIs that can be tailored to meet the Key Performance Indicators (KPIs) of various use cases, such as eMBB, mMTC, and URLLC. This ensures that each service type receives the necessary resources and quality of service, even when sharing the same physical RAN infrastructure (Fig. 10.11).

Multi-RAT Coexistence in a Single NSI One of the remarkable features of RAN slicing is the ability for radio interface protocol functions across multiple RAT to coexist within the same NSI. For example, as Fig. 10.11 shows, E-UTRA RLC and NR RLC can operate within the same NSI #1, offering greater flexibility and efficiency.

Common Protocol Functions and Schedulers RAN slicing in NG-RAN allows for the exploitation of common protocol functions or traffic schedulers across different NSI. This promotes resource sharing and operational efficiency, as a single scheduler can manage resources for multiple slices.

Flexible Radio Resource Sharing The architecture supports flexible sharing of radio resources among different NSI. This is particularly beneficial for optimizing resource utilization and accommodating diverse service requirements.

Flexible Radio Interface Protocol Stack In use cases like mMTC, some protocol layers can be deemed essential since the standalone UP can be optional. This offers a streamlined approach for services that do not require a full-fledged user plane.

By incorporating these features, RAN slicing in NG-RAN not only enhances the network's adaptability for diverse use cases but also paves the way for more efficient and flexible network operations.

10.5 5GC Architecture

10.5.1 Service-Based Architecture/SBA

5GC introduces a paradigm shift in network architecture through the adoption of Service Based Architecture (SBA), a modern approach to network design that significantly departs from classical architectures used in legacy systems like EPC.

Unlike classical architectures, which rely on dedicated point-to-point (P2P) interfaces (a.k.a. "reference points") between NF, SBA employs a more flexible, modular approach. In this architecture, as shown in Fig. 10.12, NF offer their capabilities as services that can be consumed by other NF, fostering greater flexibility, scalability, and ease of deployment. SBA is primarily applied to the control plane of the 5GC. The user plane architecture, on the other hand, remains largely similar to that of the EPC in 4G systems.

It is important to note that SBA is not mutually exclusive with classical architectures based on P2P interfaces. Both serve as perspectives to describe the interactions between 5GC NF. While SBA focuses on the services offered by NF, the classical reference point perspective emphasizes the dedicated interfaces between them, as shown in Fig. 10.13.[4] These two views coexist and can be used interchangeably, depending on specific requirements and deployment scenarios. One way to understand the structural difference between them is to consider the topologies. The traditional perspective in reference point representation resembles a mesh topology, where each pair of NF communicates directly. In contrast, the SBA can be likened to a bus topology. Here, NF expose their services on a common bus, allowing any NF to access the services of another without a direct, dedicated interface. This design not only simplifies the integration of new services but also promotes flexibility in evolving the network.

The adoption of SBA in 5GC is driven by the need to support a diverse array of use cases, from eMBB to mMTC and URLLC. Its modular and flexible design allows for rapid deployment and scaling of network services, thereby meeting the

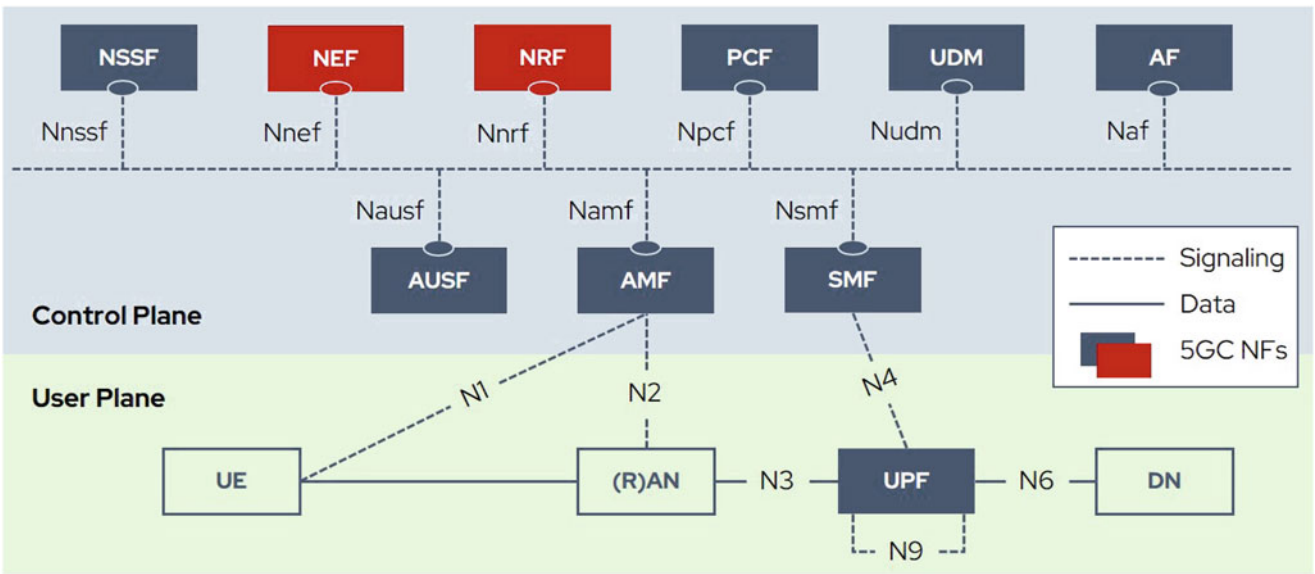


Fig. 10.12 5G system architecture visualized with service-based interfaces, 3GPP Release 15 specification (3GPP TS 23.501, 2022)

[4] Note that for the sake of clarity of the point-to-point diagrams, the NEF and NRF (highlighted in red in Fig. 10.12) are not depicted in Fig. 10.13. However, all depicted Network Functions can interact with the them as necessary.

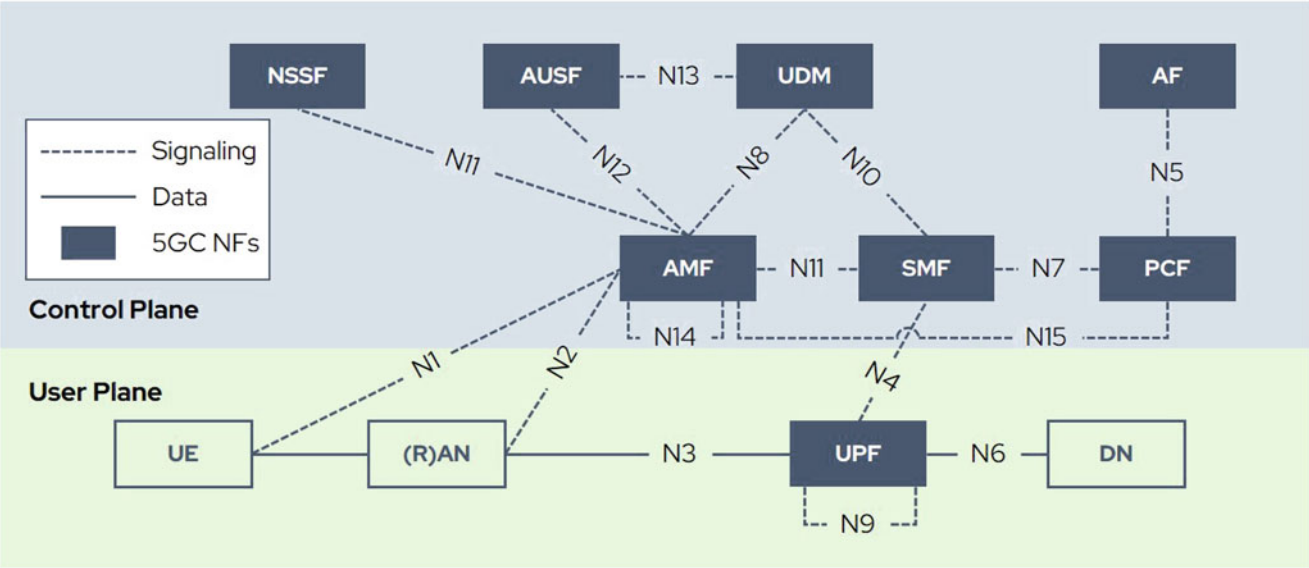


Fig. 10.13 5G system architecture visualized with in reference point representation, DN = Data Network, 3GPP Release 15 specification (3GPP TS 23.501, 2022)

diverse requirements of 5G applications. Importantly, the dynamic allocation and reallocation of network resources facilitated by SBA are especially advantageous for network slicing, a cornerstone of 5G.

One of the key advantages of SBA is its inherent support for automation and orchestration, which are vital for the efficient operation of 5G networks. Additionally, its modular design simplifies the process of incorporating new technologies and standards, making the architecture future-proof to a certain extent.

10.5.2 Enabling Technologies for SBA

The adoption of SBA in 5GC is facilitated by cloud-native design, an overarching philosophy that serves as the foundation for SBA, emphasizing modularity, scalability, and resilience. It allows 5GC to be more adaptive and responsive to the demands of various use cases. This is particularly crucial for SBA, where NF are exposed as services that must be easily consumable and independently scalable. Cloud-native design encompasses a RANge of technologies, each contributing unique advantages.

- **Virtualization:** As a subset of cloud-native design, virtualization allows NF to be abstracted from the underlying hardware, providing the flexibility to scale services up or down as needed. This dynamic resource allocation is essential for SBA, where network functions are exposed as services that must adapt to varying levels of demand.
- **Containers:** Also a feature of cloud-native design, containers offer a more lightweight solution for resource abstraction. They abstract only the application layer, making them more efficient for deploying and scaling the modular services essential in SBA.
- **Microservices:** This architectural style, integral to cloud-native design, breaks down complex NF into smaller, independent services. This modularity is key for SBA, where each NF exposes specific services that must be easily consumable and independently scalable.
- **SONs:** While not directly a part of the cloud-native or microservices paradigms, SON contribute to SBA by enabling automated configuration and optimization. This capability is particularly beneficial in a service-based environment, where the dynamic allocation and orchestration of resources are often required for optimal network performance.

Collectively, these technologies rooted in cloud-native design empower SBA to be highly adaptive, scalable, and efficient, thereby fulfilling the diverse and dynamic requirements of 5G networks.

10.5.3 5GC Network Functions

In contrast to the EPC, the 5GC represents a comprehensive evolution in its constituent NF and their interplay. The 5GC encompasses the following NF:

- **Access and Mobility Management Function (AMF)**, which *(i)* manages RAN CP interface, NAS, registration, connection, and mobility; *(ii)* provides transport for SMS between UE and respective functions; and *(iii)* supports non-3GPP access networks and may include policy-related functionalities.
- **SMF**, which *(i)* handles session management, UE IP address allocation, Dynamic Host Configuration Protocol (DHCP) functions, and UP function control; and *(ii)* manages interfaces toward policy control functions and charging data collection.
- **UPF**, which *(i)* acts as an anchor point for mobility and external PDU session interconnect; and *(ii)* manages packet routing, forwarding, inspection, and user plane policy rule enforcement.
- **Policy Control Function (PCF)**, which *(i)* governs network behavior through a unified policy framework; and *(ii)* provides policy rules to control plane functions and accesses subscription information in Unified Data Repository (UDR).
- **Network Exposure Function (NEF)**, which *(i)* exposes network capabilities and events securely; and *(ii)* translates internal-external information and may support a Packet Flow Description (PFD) function.
- **Network Repository Function (NRF)**, which *(i)* supports service discovery and maintains the NF profile of available NF instances; and *(ii)* can be deployed at different levels for network slicing and roaming contexts.
- **Unified Data Management (UDM)**, which *(i)* generates 3GPP AKA credentials and manages user identification; and *(ii)* supports access authorization, UE's serving NF registration, and SMS delivery.
- **Authentication Server Function (AUSF)**, which *(i)* supports authentication for 3GPP and untrusted non-3GPP access.
- **Non-3GPP Interworking Function (N3IWF)**, which *(i)* supports IPsec tunnel establishment with the UE for untrusted non-3GPP access; and *(ii)* manages N2 and N3 interfaces to the 5GC.
- **Application Function (AF)**, which *(i)* interacts with the 3GPP Core Network for services like traffic routing and policy control; and *(ii)* can interact directly with relevant NF or via the NEF.
- **UDR**, which *(i)* manages storage and retrieval of subscription, policy, and exposure data; and *(ii)* interacts with UDM, PCF, and NEF.
- **Unstructured Data Storage Function (UDSF)**, which *(i)* manages storage and retrieval of unstructured data by any NF.
- **Short Message Service Function (SMSF)**, which *(i)* manages SMS over NAS, including delivery, Short Message Relay Protocol (SM-RP)/Short Message Control Protocol (SM-CP) with UE, and lawful interception.
- **Network Slice Selection Function (NSSF)**, which *(i)* is responsible for selecting NSIs for the UE; *(ii)* determines allowed Network Slice Selection Assistance Information (NSSAI), configured NSSAI, and AMF set.
- **5G-EIR**, which *(i)* checks the status of Permanent Equipment Identifier (PEI) (e.g., blacklist checks).
- **Location Management Function (LMF)**, which *(i)* supports UE location determination; and *(ii)* obtains location measurements from UE and NG-RAN.
- **Security Edge Protection Proxy (SEPP)**, which *(i)* acts as a non-transparent proxy for message filtering and policing on inter-PLMN control plane interfaces; and *(ii)* provides topology hiding and security protection between Service Consumers and Service Producers.
- **Network Data Analytics Function (NWDAF)**, which *(i)* provides slice-specific network data analytics to NF; *(ii)* notifies slice-specific network status analytic information to the NF that are subscribed to it; and *(iii)* has both PCF and NSSF as primary consumers of its network analytics.
- **Charging Function (CHF)**, which *(i)* is responsible for online and offline charging mechanisms in the 5GC; *(ii)* collects real-time charging information for online charging, ensuring users have sufficient credit for service access; and *(iii)* gathers charging data for offline charging, allowing users to be billed later.

10.5.4 5GC Interfaces

In the 5GC SBA, each NF in the CP exhibits a service-based interface, over which its produced information can be consumed by any NF. The service-based interfaces are listed in Table 10.9. Meanwhile, reference points defined for the classical P2P-interface-based perspective are listed in Table 10.10.

Table 10.9 5GC NF and their exhibited service-based interfaces

NF	AMF	SMF	NEF	PCF	UDM	AF	NRF
Interface	Namf	Nsmf	Nnef	Npcf	Nudm	Naf	Nnrf
NF	NSSF	AUSF	UDR	UDSF	5G-EIR	NWDAF	CHF
Interface	Nnssf	Nausf	Nudr	Nudsf	N5g-eir	Nnwdaf	Nchf

Table 10.10 5GC reference points

N1 UE—AMF	N2 (R)AN—AMF	N3 (R)AN—UPF	N4 SMF—UPF	N5 PCF—AF	N6 (NOTE 1) UPF—DN
N7 SMF—PCF	N8 UDM—AMF	N9 UPF—UPF	N10 UDM—SMF	N11 AMF—SMF	N12 AMF—AUSF
N13 UDM—AUSF	N14 AMF—AMF	N15 (NOTE 2) PCF—AMF	N16 (NOTE 3) SMF—SMF	N17 AMF—5G-EIR	N18 Any NF—UDSF
N22 AMF—NSSF	N23 PCF—NWDAF	N24 (NOTE 4) PCF—PCF	N27 (NOTE 5) NRF—NRF	N31 (NOTE 6) NSSF—NSSF	N32 SEPP—SEPP
N33 NEF—AF	N34 NSSF—NWDAF	N35 UDM—UDR	N36 PCF—UDR	N37 NEF—UDR	N40 SMF—CHF
N50 (NOTE 7) AMF—CBCF					

Note 1: Traffic forwarding details of N6 between a UPF acting as an uplink classifier and a local data network are not specified in this Release
Note 2: In the case of a roaming scenario, between PCF in the visited network and AMF
Note 3: In roaming case between SMF in the visited network and the SMF in the home network
Note 4: Between the PCF in the visited network and the PCF in the home network
Note 5: Between NRF in the visited network and the NRF in the home network
Note 6: Between the NSSF in the visited network and the NSSF in the home network
Note 7: CBCF is an instantiation of an AF

10.6 5G Mobility Management

5G MM is a key building block of the 5GS architecture, ensuring seamless connectivity and mobility for UE. As the successor to LTE/LTE-A MM mechanisms, 5G MM introduces several innovations and enhancements to address the diverse requirements of 5G applications and services.

The 5GS architecture, while bearing similarities to the EPS, introduces several enhancements. Mobility management procedures, such as the establishment of connectivity, network discovery, and selection procedures for 5GS, are largely retained from EPS. However, the 5GS introduces specific enhancements and new features.

10.6.1 Architectural Evolution: AMF and SMF

In 5GS, the functionalities that were previously executed by MME in EPS are distributed between two NF: the AMF and the SMF. This separation of 5G MM (via N2 interface) from session management (via N4 interface) ensures that only necessary signaling is exchanged between the UE, NG-RAN, and 5GC during mobility events and therewith offers potential advantages in terms of scalability, flexibility, and efficiency.

10.6.2 New MM Procedures

The MM in 5GS, as illustrated in Fig. 10.14, is generally based on the procedures of registration/deregistration, which replaces the attach/detach and TA update procedures of EPS. Unlike the attach procedure, the registration procedure in 5GS is more versatile. It allows a UE to register its presence in the network, request specific services, and establish user plane resources. The procedure can be initiated for various reasons, including periodic registration updating, mobility, or service requests. The UE can also indicate its preference for a specific network slice during this procedure.

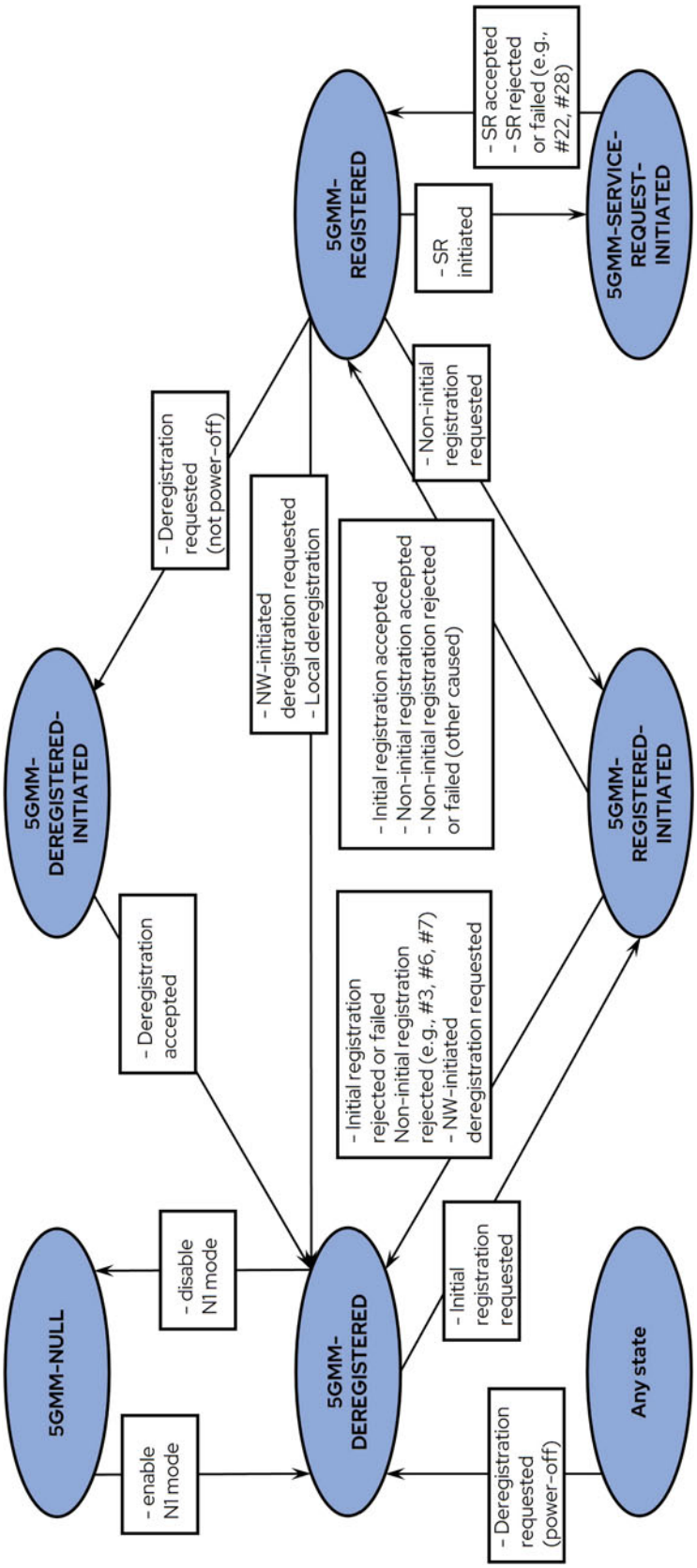


Fig. 10.14 The 5G MM state machine model

10.6.3 Energy Efficiency Enhancements

The Mobile Initiated Connection Only (MICO) mode, a.k.a. the minimum control signaling, is introduced in 5GS to reduce the frequency of signaling between the UE and the network. In MICO mode, the UE reduces the frequency of registration updates, especially when it is in a stationary state or when data transmission is not expected for a prolonged period. A UE enters MICO mode based on network instructions and exits when there is a need for data transmission or upon receiving specific network instructions. This mode is particularly beneficial for UE that only occasionally transmit data.

The RRC Inactive state, which we have introduced in Sec. 10.3.4, allows the UE to quickly resume communication without undergoing the full re-establishment procedure, streamlining the transition between inactive and active states.

10.6.4 Service Area & Mobility Restrictions, and N2 Management

In the 5GS, mobility management is enhanced with tailored restrictions to ensure efficient access and service delivery for UEs. The *Service Area Restriction* delineates areas where the UE can initiate communication. Specifically, within an *Allowed Area*, the UE freely communicates with the network, while in a *Non-Allowed Area*, the UE is restricted from initiating user services, though it can still perform mobility-related signaling and respond to network-initiated messages.

Beyond geographical restrictions, 5GS introduces other mobility constraints. These include RAT restrictions, which specify the 3GPP RATs a UE can access, and CN type restrictions, determining if a UE can access the 5GC, EPC, or both. Additionally, there are *Forbidden Areas* where the UE cannot initiate any network communication.

The efficient execution of these restrictions is facilitated through the N2 interface, which bridges the AMF and the NG-RAN (see Table 10.10). N2 handles the signaling related to mobility management, session management, and more. In the context of 5G MM, N2 management plays a pivotal role in handling UE context, paging, handovers, and other mobility-related procedures. The design of N2 management in Release 15 is geared toward ensuring efficient signaling, reduced latency, and seamless mobility.

10.6.5 Control of Overload and Unified Access Control

To manage the load produced by UE toward the 5GS, the system introduces mechanisms like dynamic adjustment of update intervals, prioritizing UE based on QoS requirements, and temporarily barring certain UE or services.

The Unified Access Control (UAC) in 5GS is an evolution of mechanisms like Access Class Barring (ACB) from EPS.While ACB in EPS was primarily used to prevent congestion during mass events or emergencies by barring certain access classes, UAC in 5G provides a more granular approach. It considers factors like the UE's subscription profile, current network conditions, and even specific services or applications. The introduction of UAC impacts 5G MM by providing a more flexible and efficient way to control UE access to the network, ensuring better resource utilization and improved user experience.

10.6.6 Interworking with EPC: A More Streamlined Approach

Interworking between 5GS and EPC in terms of handovers is more streamlined compared to the interworking between LTE/LTE-A and UMTS. This is primarily because both 5G and LTE/LTE-A are purely PS networks. The introduction of EN-DC in 5G allows UE to maintain simultaneous connections to both E-UTRAN and NG-RAN, ensuring seamless handovers and improved data rates.

10.7 5G Security

10.7.1 Overview

The advent of 5G has marked a transformative era in mobile communication, offering a vast suite of capabilities and services. However, with these advancements come intricate security challenges. While 5G inherits many security principles from 4G, it introduces several enhancements to address the unique challenges.

5G is poised to serve a diverse RANge of end-devices, from traditional mobile broadband devices to connected appliances and industrial applications. This extensive integration accentuates the significance of privacy. To address this, 5G ensures that identifiers, suchlike the Subscription Permanent Identifier (SUPI), are never transmitted in clear text over the air.

The intricate architecture of 5G brings forth challenges in ensuring user privacy, message confidentiality, and protection against cyber-attacks. The introduction of diverse use cases highlights the potential severe repercussions of successful attacks, especially in critical infrastructures. Guidelines from the *German Federal Network Agency* emphasize the need for secure network operations, including encryption techniques, protection against various attacks, and disabling unused services. Furthermore, 5G's landscape is characterized by multiple stakeholders, from transport infrastructure providers to third-party industries integrated with their virtual networks or MEC applications. This vast ecosystem presents a wide variety of system components, subsystems, infrastructures, platforms, and functions that must be considered in the requirements and considerations for security.

To address these challenges and ensure its security against such diverse risks, 5G demands protection across multiple layers. Building on 4G foundations, 5G introduces further enhancements like optional integrity protection of UP traffic, especially vital for IoT scenarios vulnerable to unauthorized alterations. Generally, 5G security is anchored in a three-pillar model (Schneider & Urban, 2020): the security of communication network, the security of cloud infrastructure, and the standardized 3GPP security architecture.

10.7.2 3GPP Security Architecture for 5G

The 3GPP security architecture for 5G, which is illustrated in Fig. 10.15, forms a cornerstone of the three-pillar model of 5G security. This architecture is meticulously designed to address the unique challenges and requirements of 5G, ensuring robustness, flexibility, and adaptability in the face of evolving threats.

Central to the 3GPP security architecture are the six *security domains*. These domains encompass various facets of security, ensuring a holistic approach:

- **Network access security:** Focuses on the protection of communication between the UE and the 5G network, ensuring confidentiality, integrity, and authentication.
- **Network domain security:** Addresses the security of intra-operator and inter-operator communication, safeguarding data and signaling traffic.
- **User domain security:** Concentrates on securing access to mobile equipment, primarily through PIN and passwords, ensuring that unauthorized users cannot access a UE.
- **Application domain security:** Pertains to the security of end-to-end communication at the application layer, ensuring protection against eavesdropping and tampering.
- **SBA domain security:** Specifices to the 5G core network, ensuring the security of communication between different network functions in the service-based architecture.
- **Visibility and configurability of security:** Enables the user to check the security level of their current communication and, to some extent, influence it.

The security requirements for 5G are stringent and comprehensive. They encompass protection against bidding-down attacks, robust authentication and authorization procedures, and provisions for emergency calls. Each network element, from the UE to the various core network functions, has specific security requirements tailored to its role and potential vulnerabilities. These requirements, as listed in Table 10.11, ensure that the 5G network remains resilient against a wide array of threats, from eavesdropping to more sophisticated cyber-attacks.

Table 10.11 Security requirements for 5G network elements and 5G network functions (3GPP TS 23.501, 2023)

5G network elements or 5G NF	Security requirements
UE	• Encryption of signaling and user data between UE and gNodeB for reasons of confidentiality • Ensuring data integrity for signaling and user data between UE and gNodeB • Secure storage and processing of the login information from the user profile • Protection of privacy through encryption and secure storage of keys in the USIM • Calculation of the Subscription Concealed Identifier (SUCI)
gNodeB	• Encryption of signaling and user data between UE and gNodeB for reasons of confidentiality • Ensuring data integrity for signaling and user data between UE and gNodeB • Authenticating and authorizing a gNodeB during setup and configuration • Protection of the gNodeB software • Protection of the keys used and stored in the gNodeB • Secure processing and storage of user and signaling data • Providing a secure environment for all sensitive data • Secured transmission on F1 interface when splitting a gNodeB into CU and DU • Secure transmission on E1 interface when dividing a CU into CU-CP and CU-UP
AMF	• Because of the confidentiality encryption of NAS signaling • Ensuring data integrity for NAS signaling • Triggers the primary authentication of the UE via SUCI
UDM	• Long-term keys for authentication and security association must be protected and must not leave the UDM/Authentication Credential Repository and Processing Function (ARPF) environment. • Provides Subscription Identifier Deconcealing Function (SIDF) service
SIDF	• Responsible for resolving the SUPI from the SUCI
AUSF	• Processes authentication requests for 3GPP and non-3GPP access • Informs about it UDM • Transfers the SUPI to the visited PLMN (VPLMN) after successful authentication
5GC in general	• Creation of trust zones, in any case between different providers • Secure discovery and registration of NF in the SBA • Authentication between NF producer and NF consumer • Validation of each received message by NF • Secure end-to-end connections for the application layer between 5GC.
NRF	• The NRF and service-requesting NF must authenticate each other. • The NRF provides authentication and authorization to NF for secure communication between themselves.
NEF	• Ensures confidentiality and data integrity between NEF and AF • Mutual authentication • Does not communicate information about network slice or SUPI to the exterior
SEPP	• Protects communication between NF in different PLMN • Mutual authentication with corresponding SEPP • Hiding of the own SBA (topology hiding) • Application Layer Gateway functionality

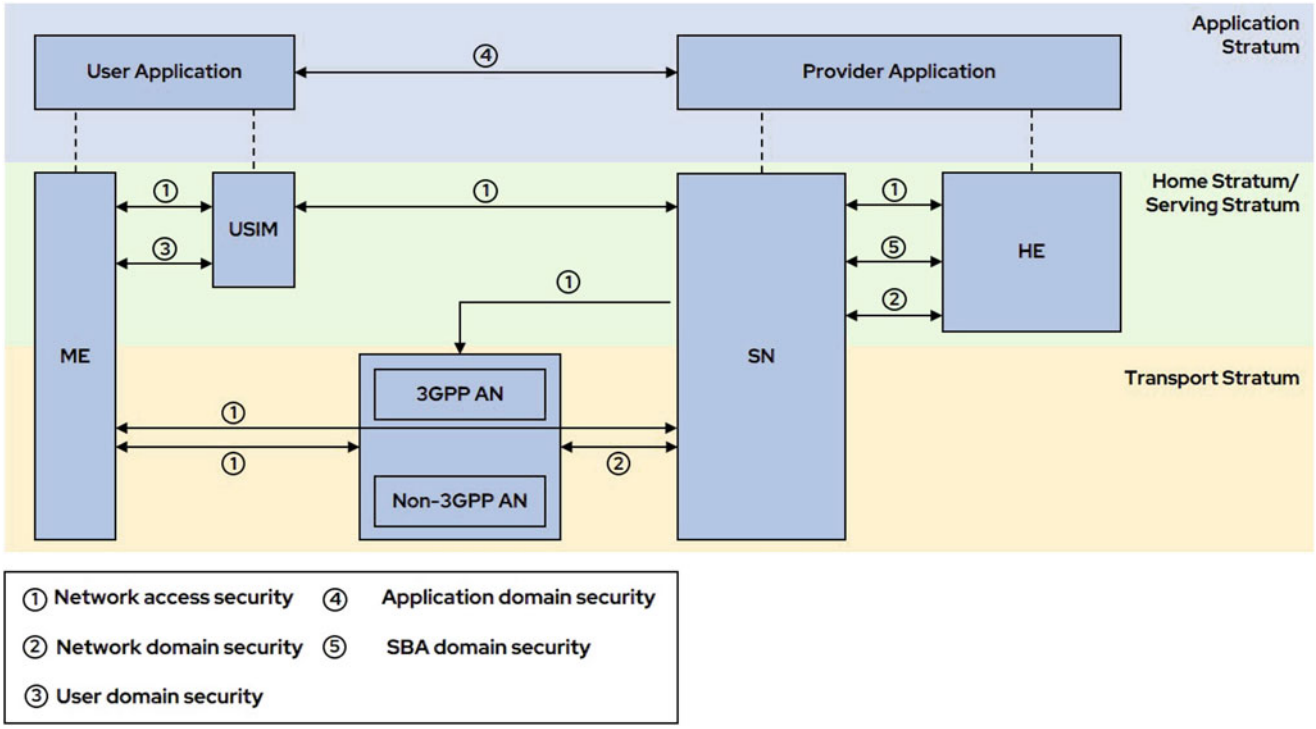


Fig. 10.15 Overview to the 3GPP security architecture (3GPP TS 23.501, 2023)

10.7.3 NG-RAN Enhancements for Security

The NG-RAN stands central to the 5G ecosystem, ensuring seamless connectivity and robust security. The enhancements introduced in NG-RAN are meticulously designed to address the specific security requirements of both the UE and the gNodeB, as detailed in Table 10.11.

For the UE, the NG-RAN ensures that both signaling and user data between the UE and the gNodeB are encrypted, safeguarding the confidentiality of the transmitted information. Alongside encryption, the integrity of signaling and user data is ensured, preventing potential tampering or unauthorized alterations during transmission. The UE is equipped with mechanisms that allow for the secure storage and processing of login information from the user profile. This ensures that only authorized users can access the network and utilize its services. Furthermore, the UE employs encryption techniques and securely stores keys within the USIM, ensuring user privacy. This is further bolstered by the UE's capability to calculate the SUCI, which aids in preventing potential identity theft or unauthorized tracking.

On the other hand, the gNodeB ensures that all signaling and user data between the UE and itself are encrypted. This not only ensures confidentiality but also guarantees the integrity of the data, ensuring that it remains unaltered during transmission. The gNodeB undergoes a rigorous authentication and authorization process during its setup and configuration. This ensures that only legitimate base stations can join the network and serve users. The gNodeB's software is shielded against potential threats, ensuring its uninterrupted operation. Additionally, the keys used and stored within the gNodeB are protected, preventing unauthorized access or potential breaches. The gNodeB is designed to securely process and store user and signaling data. It provides a secure environment for all sensitive data, ensuring that it remains shielded from potential threats. The modularization of the gNodeB into the CU and DU necessitates secure communication between these components. The NG-RAN ensures secured transmission on the F1 interface during this split. Furthermore, when the CU is divided into CU-CP and CU-UP, the E1 interface ensures secure transmission.

10.7.4 5GC Security Enhancements

Serving as the backbone of the 5G network, the 5GC is entrusted with the task of coordinating various network functions and facilitating seamless communication. Its central position underscores the imperative nature of its security. The intricate security requirements tailored for 5GC, as delineated in Table 10.11, underscore the comprehensive measures taken to shield the network from potential adversarial actions.

General Security Measures in 5GC The architecture of 5GC emphasizes the creation of trust zones, particularly when interfacing with different providers. This strategic design promotes the secure discovery and registration of NF within the SBA. Authentication is a standard procedure for every NF in the 5GC, be it a producer or consumer, fortifying the communication environment. Additionally, each message received by NF is subjected to rigorous validation, and end-to-end connections are fortified for the application layer between 5GC.

Emphasis on Secure Storage and Processing of Subscription Credentials and Keys The UDM is instrumental in safeguarding long-term keys used for authentication and security association. The AUSF, on the other hand, is adept at handling authentication requests, catering to both 3GPP and non-3GPP access.

Ciphering and Integrity Protection Mechanisms in 5GC Through the AMF, NAS signaling undergoes encryption, ensuring data remains confidential. Concurrently, the NEF establishes a secure bridge between itself and the AF, preserving both confidentiality and data integrity.

Strategies for Upholding Privacy in 5GC The SIDF takes the lead in deciphering the SUPI from the SUCI, a crucial step in maintaining user privacy. The AUSF further augments privacy by transmitting the SUPI to the visited PLMN post successful authentication, while the NEF remains vigilant, preventing any external communication of network slice or SUPI details.

Inter-PLMN Security Facilitated by 5GC The SEPP stands as a guardian, ensuring protected communication between NF across different PLMN. In tandem, the NRF mandates mutual authentication between itself and the service-requesting NF.

10.8 Summary

The chapter begins by elucidating the foundational concepts and terminologies associated with 5G, setting the stage for a deeper exploration of its intricate architecture. The RAN evolution of is detailed, emphasizing the innovations in the NR and the enhancements it brings to the table. The chapter also sheds light on the 5GC and its service-based architecture, highlighting its flexibility and adaptability. The role of network slicing in ensuring efficient resource utilization and tailored services is discussed, showcasing its potential in revolutionizing network operations. The chapter culminates with an in-depth examination of the security enhancements in 5G, underscoring the importance of safeguarding this next-generation network against a myriad of threats. NG-RAN and 5GC, introduced in 3GPP Release 15, mark a significant step forward in mobile communication. Building on the foundational concepts and key technologies of previous generations, 5G initiates a paradigm shift by connecting not only people but also a wide array of devices, enabling new applications and services. As outlined in Chap. 9, the transition to 5G was driven by the need for higher data rates, lower latency, and the capability to connect a vast number of devices. However, the substantial investment required for a full 5G deployment posed challenges for operators. The NSA deployment option emerged as a practical solution, allowing for a more gradual transition by utilizing existing LTE infrastructure. This approach facilitated a quicker initial deployment, enabling operators to offer 5G services while continuing to develop the full NG-RAN and 5GC infrastructure.

10.9 Exercises

1. Compare and contrast the SA and NSA deployment options for 5G networks. Discuss the implications of each on network transition strategies.
2. Explain the concept of network slicing in 5G. How does it enable the support of diverse service requirements?
3. Describe the differences in waveform and radio frame design between 5G NR and LTE-A. Discuss the impact of these differences on network efficiency and user experience. Additionally, explain how the radio interface protocol design has evolved from LTE-A to 5G NR.
4. Sketch the NG-RAN architecture, highlighting the roles of the CU and the DU. How does RAN slicing contribute to network flexibility?
5. Outline the SBA of 5GC. How do NF interact in this architecture, and what are the benefits of such an arrangement?
6. Elucidate the role of the AMF and the SMF in 5G mobility management. Describe the key mobility management procedures that are facilitated by these entities in the 5G network.
7. Detail the specific security protocols and features introduced in 5G to enhance user privacy and network integrity. Discuss the roles of SEPP, UDM, and the AUSF in the 5G security framework. How do these components work together to mitigate common security threats?

11 Evolution Toward Sixth-Generation (6G) Mobile Cellular Systems

11.1 The Race on Developing 6G

In April 2019, the era of 5G mobile communications commenced as South Korea's three mobile operators, namely SK Telecom, LG U+, and KT, engaged in a competitive race with Verizon, a US carrier, to offer the world's first 5G commercial service. Over the past several years, we have witnessed a significant leap in global 5G penetration, along with a remarkable surge in 5G subscriptions across the world. According to the statistics from the GSM Association (GSMA), there were already more than one billion 5G subscribers by the end of 2022, and this figure is projected to surpass 4.6 billion worldwide in 2028, constituting over 50 percent of all mobile subscribers. Currently, there are over 200 operational 5G networks globally, and over 700 different 5G smartphone models are available for users. By 2028, 5G is anticipated to become the prevailing mobile access technology in terms of subscriptions. For instance, China has achieved a 90 percent coverage of 5G with the installation of approximately 3 million 5G base stations.

Unlike previous generations of cellular networks that focused primarily on human-centric communication services, 5G has a much wider range of applications. It not only aims to provide mobile broadband but also caters to massive machine-type communications and ultra-reliable, low-latency communications for mission-critical uses. The introduction of 5G expands the sphere of mobile communications from human-centric to the Internet of Everything and meanwhile from the consumer market to vertical industries. Accordingly, the scale of mobile subscriptions has expanded significantly, not only connecting billions of people but also creating countless interconnections among humans, machines, and things (Andrews et al., 2014). In addition to the billions of human consumers, it is projected that the number of connected IoT devices will reach 25 billion by 2025. In the year 2020, the outbreak of the COVID-19 pandemic resulted in a devastating loss of human life worldwide and posed significant challenges to society and the economy. However, this public health crisis underscored the crucial role of networks and digital infrastructure in maintaining societal functions and keeping families connected. 5G-enabled applications, such as remote surgery, online education, high-definition video conferencing, autonomous vehicles, unmanned delivery, robots, contactless healthcare, and automated manufacturing, demonstrated their unique value during the fight against the pandemic.

Despite the ongoing global deployment of 5G mobile networks, both academia and industry have already shifted their focus on the next-generation cellular technology known as 6G or International Mobile Telecommunications-2030 (IMT-2030). While there were debates within the wireless community regarding the necessity of developing 6G and whether the mobile generation count should stop at 5 (Fitzek & Seeling, 2020), a collection of research groups, standardization bodies, regulatory organizations, and government agencies have initiated various programs to discuss the vision of 6G and develop key technologies. In July 2018, the *International Telecommunication Union, Telecommunication Standardization Sector* (ITU-T) established a focus group called *Technologies for Network 2030*. This group aimed to explore the capabilities of networks for the year 2030 and beyond, as well as the innovative scenarios that lie ahead. According to its report (ITU-T NET-2030, 2019), 6G is expected to support disruptive applications such as virtual reality (VR), augmented reality (AR), and mixed reality (MR), which fall under the umbrella term of extended reality (XR), the Internet of Everything, Industry 4.0, connected and autonomous vehicles, as well as yet-to-be-conceived use cases like the Metaverse, holographic-type telepresence, the Tactile Internet (Fettweis et al., 2014), digital twins, full immersiveness, multi-sense experiences, and blockchain technology.

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5_11

In the year 2020, the European Commission (EC) started to sponsor beyond 5G research in Europe through its calls under Horizon 2020, namely ICT-20: *5G Long-Term Evolution* and ICT-52: *Smart Connectivity beyond 5G*, which have initiated a batch of pioneering research projects aimed at exploring potential beyond 5G technologies. At the beginning of 2021, the EC sponsored the European 6G flagship research project "*Hexa-X: A holistic flagship towards the 6G network platform and system, to inspire digital transformation, for the world to act together in meeting needs in society and ecosystems with novel 6G services*," followed by the second phase of European level 6G research *Hexa-X-II* in early 2023. In addition, the EC has announced its *Gigabit Connectivity* strategy to accelerate investments in both 5G and 6G, with the aim of shaping Europe's digital future (EU Gigabit Connectivity, 2020). In October 2020, the Next Generation Mobile Networks (NGMN) alliance launched its *6G Vision and Drivers* project (NGMN, 2021), aiming to provide early guidance and direction for global 6G activities.

At its meeting in February 2020, the ITU-R Working Party 5D decided to start the work plan for a report on future technology trends for the future evolution of terrestrial IMT systems (ITU-R WP5D, 2020). After around two and a half years, the ITU-R WP 5D published its outcomes in the report ITU-R M.2516 (2022) "*Future technology trends of terrestrial International Mobile Telecommunications systems towards 2030 and beyond*." It elaborates on technology enablers for enhancing radio interface and radio access networks, as well as emerging technology trends for enabling new functions and services. In early 2021, the ITU-R WP 5D officially launched the study of a new recommendation M. [IMT Framework for 2030 and Beyond], commonly known as the 6G vision. After continuous discussion and a total of 156 contributions, the draft of the new recommendation for IMT-2030 vision was completed on schedule and finalized on June 22, 2023 at the 44th ITU-R WP 5D meeting in Geneva. As the first and fundamental milestone of developing 6G, this recommendation specifies the framework and overall objectives of International Mobile Telecommunications in 2030 and beyond.

Motivated by the revolutionary force of 5G, the governments of many countries recognized the significance of mobile communications technologies for driving economic prosperity and sustainable growth. In the past few years, many countries have set up research initiatives officially for the development of 6G technology. The world's first 6G effort, "*6G-Enabled Wireless Smart Society and Ecosystem (6Genesis)*," renamed "*6G Flagship Program*" later, was carried out by the University of Oulu in April 2018, as part of the Academy of Finland's flagship program. This project focused on groundbreaking 6G research, with four interrelated strategic areas including wireless connectivity, distributed computing, devices and circuit technology, and services and applications. In September 2019, the world's first 6G white paper "*key drivers and research challenges for 6G ubiquitous wireless intelligence*" was published as an outcome of the first 6G Wireless Summit (6G Flagship Program, 2019). Subsequently, a series of white papers have been published, covering 12 specific areas of interest, such as machine learning, edge intelligence, localization, sensing, and security.

The establishment of the "*Next G Alliance*" in October 2020 by the ATIS marked an industry-led initiative with the goal of advancing North American mobile technology leadership in the field of 6G over this decade. Tech giants such as AT&T, T-Mobile, Verizon, Qualcomm, Ericsson, Nokia, Apple, Google, Facebook, and Microsoft are among the founding members of this initiative. The Next G Alliance places significant emphasis on the commercialization of technology and aims to encompass all aspects of 6G, including research, development, manufacturing, standardization, and market readiness. Additionally, in 2015, SpaceX, a notable US company known for its reusable rockets, announced the Starlink project. This project aims to deploy a large-scale low Earth orbit (LEO) communications satellite constellation to provide ubiquitous Internet access services worldwide. The FCC has approved the initial plan to launch 12,000 satellites, and an application for an additional 30,000 satellites is currently being considered. While it would be an exaggeration to state that Starlink will replace 5G or be considered 6G, the impact of such a large-scale LEO satellite constellation on 6G and future mobile industry developments should be taken into serious consideration.

The Chinese Ministry of Science and Technology, in collaboration with five other ministries or national institutions, initiated the research and development efforts for 6G technology in November 2019. This milestone also marked the establishment of a working group called the *IMT-2030(6G) Promotion Group*, which was tasked with managing and coordinating the program. Additionally, an expert group consisting of 37 top researchers from academia, research institutes, and industry was formed. In June 2021, the IMT-2030(6G) Promotion Group released a white paper (IMT-2030(6G) Promotion Group, 2021), entitled "*6G Vision and Candidate Technologies*." The white paper provides an overview of the group's cutting-edge research findings, including the vision for 6G, the driving forces behind its development, potential use cases, ten candidate technologies, and additional insights.

In the latter part of 2017, the Japanese Ministry of Internal Affairs and Communications established a working group to investigate next-generation wireless technologies. The research findings of this group highlighted that 6G technology should provide transmission rates at least ten times faster than 5G, near-instant connectivity, and the ability to connect a massive number of devices, up to 10 million per square kilometer. In December 2020, Japan launched the *Beyond 5G Promotion*

Consortium (B5GPC) with the aim of accelerating the development of 6G while enhancing the country's international competitiveness through collaboration between industry, academia, and government. B5GPC published its first white paper titled "*Beyond 5G white paper: Message to the 2030s*" in March 2022 (Beyond 5G Promotion Consortium, 2022). This white paper summarizes the requirements and expectations of various industries for 6G and outlines the necessary capabilities and technological trends. Moreover, South Korea has announced its ambition to establish the world's first 6G trial by 2026. They have also introduced the *K-Network 2030* initiative, which seeks to support the development of crucial 6G technologies. This initiative includes objectives such as the development of cloud-native networks using domestically made AI chips, the launch of a low-orbit communications satellite by 2027, and the establishment of an open RAN ecosystem for domestic companies.

In February 2021, the German Federal Ministry of Education and Research (BMBF) introduced a funding program named "6G Vision" as part of Germany's broader initiative to establish itself as a leader in 6G technology. In August 2021, four 6G research hubs, including 6G-life, 6GEM, 6G RIC, and Open6GHub, were established under the umbrella organization of the 6G platform (Fitzek et al., 2022). The allocated budget for these endeavors amounted to around 250 million euros, supporting 160 research groups across 21 universities, 15 research institutes, and over 40 small and medium enterprises. In the following year, 18 6G industry projects, such as 6G-ANNA, 6G-TakeOff, 6G-Terafactory, and 6G-Next, along with seven resilience-focused projects like HealthNet, AKITA, and ConnRAD, were initiated (Schotten, 2023). In France, the National Agency for Research (ANR) launched a national acceleration strategy in 2021, known as the *France 2030 Plan*, which specifically targets future communication technologies, digital transition, telecommunications, and global innovation. On May 1, 2023, as part of the *France 2030 Plan*, the Priority Research Programs and Equipment (PEPR) 5G and future networks program commenced with the objective of developing advanced technologies for 5G and future networks. This project aims to integrate considerations of environmental and societal impacts, as well as data security. The ANR PEPR 5G program comprises ten interlinked projects, fostering collaboration and synergy.

11.2 Mobile Traffic Growth by 2030

The driving force behind the next-generation cellular system stems from several factors. Firstly, there is the continuous growth of mobile traffic and mobile subscriptions, accompanied by the emergence of disruptive services and applications on the horizon. Additionally, the inherent demand within the mobile communication society to enhance network efficiencies – encompassing cost efficiency, energy efficiency, spectrum efficiency, and operational efficiency – plays a crucial role in shaping this system's development. Moreover, technological advancements like native AI, integrated sensing and communication, communication-computing convergence, new networking paradigm, and large-scale satellite constellations, present an opportunity for cellular networks to evolve into a more capable, robust, and efficient system. This evolution not only aims to meet the current service requirements but also holds the potential to introduce entirely novel and groundbreaking services. In this section, we focus on the first driving factor of 6G, i.e., the trend of mobile traffic, which is anticipated to maintain its explosive growth until 2030. The next section sheds light on disruptive use cases and applications.

We find ourselves in an extraordinary era characterized by the rapid emergence and evolution of numerous interactive, intelligent products, services, and applications, placing immense pressure on mobile communications. Looking ahead, it is evident that the existing 5G system will face challenges in handling the overwhelming volume of mobile traffic expected in 2030 and beyond. Back in 2015, the ITU-R published a report (ITU-R M.2370, 2015) that analyzed the growth of IMT traffic from 2020 to 2030. According to this report, the primary drivers behind the projected surge in traffic are:

- *Proliferation of Video Services*: The utilization of video-on-demand services is expected to keep expanding, accompanied by a continuous increase in video resolution. People's desire to access high-quality visual content remains strong, irrespective of the delivery method. According to a study conducted by Bell Labs, video streaming already constituted nearly two-thirds of all mobile traffic in 2016.
- *Terminal Diversification*: By 2022, global smartphone ownership has surpassed 5 billion, and over 1 billion new smartphones are sold each year, accounting for approximately 70% of the world's population. So far, human users have been the primary consumers of data. However, there has been a notable emergence of intelligent smart machines, wearable electronics, and VR glasses, which are also significant consumers of data. Furthermore, it is anticipated that a variety of wearable devices like skin-patches, bio-implants, and exoskeletons could be combined with cutting-edge man–machine interfaces, such as gestures, haptics, and brain sensors, fostering the development of new applications. While smartphones continue to be prevalent, non-portable terminals like smart automobiles, unmanned aerial vehicles (UAVs), vessels, and

robots equipped with advanced multisensory integration and intelligent capabilities are projected to play an increasingly important role in various fields of future society.

- *Mobile APP's Uptake*: The proliferation of mobile APPs is rapidly increasing. In 2013, the annual global APP downloads reached 102 billion, and by 2017, this number had surged to 270 billion. However, a significant portion of these downloaded APPs tends to be used only once and then discarded. This influx of APP downloads and their usage patterns significantly contribute to the surge in mobile broadband traffic. Moreover, the continuous updates required for the hundreds of billions of applications also add to the increased mobile broadband traffic.

Apart from the prominent factors mentioned earlier that drive data traffic growth, several other characteristics and trends are poised to impact the overall traffic demand by the year 2030.

- *Wide Deployment of 5G Networks*: New technologies will improve the perceived quality of experience and decrease the cost per bit, which in turn creates more traffic demand. In addition, 5G expands the sphere of mobile communications from human-centric to M2M and meanwhile from the consumer market to vertical industries. Accordingly, the scale of mobile subscriptions has expanded significantly, not only connecting billions of people but also creating countless interconnections among humans, machines, and things.
- *Internet of Things*: M2M applications and IoT devices constitute one of the most rapidly expanding segments in the realm of mobile services, ultimately leading to an escalation in mobile data demand. The sheer volume of M2M connections is projected to be several orders of magnitude greater than the world's population. As a result, billions of IoT devices have the potential to utilize cellular networks for accessing online services and establishing connections with each other. This exponential growth in M2M connectivity will contribute significantly to the overall demand for mobile data in the coming years.
- *Enhanced Screen Resolution*: Continuous improvement in the screen capabilities, e.g., 4K Ultra-High-Definition (UHD), and increasing demand for video downloading and streaming will bring more traffic on mobile networks.
- *Proliferation of Ambient Screens* or info-bearing surfaces to Internet-connected devices for up-to-date information, such as screens in elevators and buses, will increase traffic.
- *Cloud Computing*: The demand for mobile cloud services is expected to grow because users are increasingly adopting more services that are required to be ubiquitously accessible. With the increasing number of users connecting through the mobile network to the cloud, the mobile data traffic between mobile terminals, cloud servers, and cloud storage will continue to grow.
- *Fixed Broadband Replacement*: In areas and contexts where eMBB is used as an alternative to wired broadband, such as copper, cable, and optic fiber, this would contribute to an increase in IMT traffic.
- *Multimedia Streaming*: The usage of mobile devices for multimedia streaming entertainment has increased significantly, driven by various factors such as time shift (enabled by cloud expansion), space shift (availability of content anywhere), and device shift (multi-screen capability and seamless switching between mobile and portable devices). In 2014, live TV still dominated the audio-visual services worldwide, accounting for 90% of the market and growing by 4.2% compared to the previous year. Meanwhile, over-the-top video transmission represented 4.4% of the market but experienced remarkable growth of 37%. Currently, a significant portion of audio-visual traffic is delivered through non-IMT networks, but there is a trend to transfer this traffic into IMT networks.

In a nutshell, the traffic over mobile networks will continuously grow in an explosive manner due to the proliferation of rich-video applications, terminal diversification, enhanced screen resolution, IoT services, mobile cloud services, etc. According to the estimation by ITU-R M.2370 (2015), the global mobile traffic will reach up to 5016 EB per month in 2030 compared with 62 EB in 2020, where 1 exabyte (EB) equals 1,000,000,000 gigabytes (GB). A report from Ericsson (Ericsson Report, 2020) reveals that the global mobile traffic has reached 33 EB per month at the end of 2019, which justified the correctness of ITU-R's estimation. Over the past decade, there has been a remarkable surge in the usage of smartphones and tablets, largely driven by the widespread availability of mobile broadband. This exponential growth is expected to persist throughout the 2020s as smartphones and tablets have not yet reached their full market saturation, particularly in developing nations and remote areas. Additionally, the emergence of novel user terminals like wearable electronics, IoT devices, and XR glasses is happening at an unprecedented speed, with consumers adopting these technologies rapidly. As a result, it is expected that the total number of eMBB subscribers worldwide will reach 17.1 billion by 2030, as shown in Fig. 11.1.

On the other hand, the traffic demand per eMBB user continuously rises, in addition to the rising number of eMBB users. That is mainly because of the popularity of mobile video services, such as YouTube, Netflix, and Tik-Tok, and the stable improvement of screen resolution on mobile devices. The traffic coming from mobile video services already accounts for two-thirds of all mobile traffic nowadays (Ericsson Report, 2020) and is estimated to be more dominant in the future. In

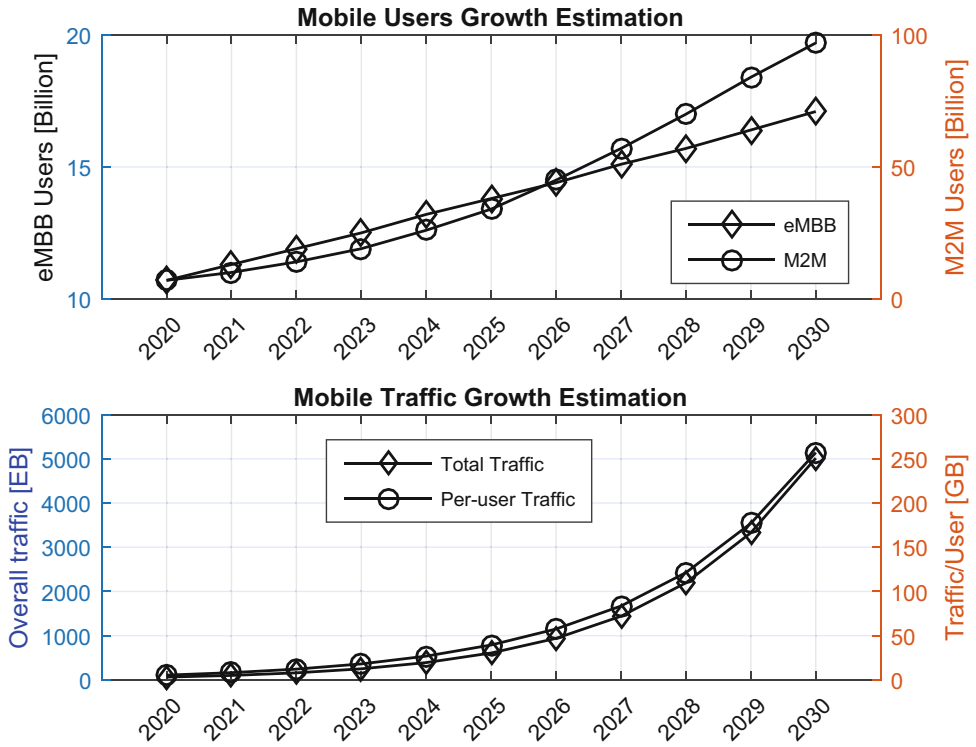


Fig. 11.1 The prediction of global mobile subscriptions and traffic growth until 2030. Source: (Jiang et al., 2021)

some developed countries, strong traffic growth before 2025 will be driven by rich-video services, and a long-term growth wave will continue due to the penetration of AR and VR applications. The average data consumption for every mobile user per month, as illustrated in Fig. 11.1, will increase from around 5 GB in 2020 to over 250 GB in 2030. In addition to human-centric communications, the scale of M2M connections will increase more rapidly and become saturated no earlier than 2030. The number of M2M subscriptions will reach 97 billion, around 14 times over 2020 (ITU-R M.2370, 2015), serving as another driving force for the explosive growth of mobile traffic.

11.3 Potential Services and Applications of IMT-2030

The role of IMT-2030 in the future will be to establish intelligent connections between numerous devices, processes, and humans, forming a global information grid. This advancement will create fresh opportunities across various industries and sectors. Given their distinct development cycles, a wide array of potential advancements and vertical transformations will continue to unfold in the post-2030 era. The demand for higher data rates will persist leading up to 2030, with peak data rates potentially reaching the order of magnitude of terabits per second in indoor environments. It will necessitate extremely wide bandwidths, paving the way for THz or sub-THz communications. Concurrently, a substantial portion of vertical data traffic will be measurement-based or actuation-related small packets. Meeting the requirements for such scenarios will demand extremely low latency in tight control loops, which, in turn, may require short over-the-air latencies to allow time for computation and decision-making processes. Furthermore, the reliability and QoS requirements in various vertical applications will increase to ensure that required services are readily available in their specific areas of need. For instance, industrial devices, processes, and future haptic applications, including multi-stream holographic applications, will demand strict timing synchronization with tight requirements for jitter. This emphasizes the need for precise and reliable connectivity solutions in the evolving IMT landscape.

In its recent report ITU-R M.2516 (2022), the ITU-R envisions new services and application trends for IMT toward 2030 and beyond. That is,

- Networks will support enabling services that help to steer communities and countries toward reaching the United Nations' sustainable development goals.
- Customization of user experience will increase with the help of user-centric resource orchestration models.
- Localized demand–supply–consumption models will become prominent at a global level.
- Community-driven networks and public–private partnerships will bring about new models for future service provisioning.
- Networks will have a strong role in various vertical and industrial contexts.
- Market entry barriers will be lowered by the decoupling of technology platforms, enabling for multiple entities to contribute to innovations.
- Empowering citizens as knowledge producers, users, and developers will contribute to a process of human-centric innovation, contributing to pluralism and increased diversity.
- Privacy will be strongly influenced by increased platform data economy or sharing the economy, emergence of intelligent assistants, connected living in smart cities, transhumanism, and digital twins.
- Monitoring and steering of circular economy will be possible, helping to create a better understanding of sustainable data economy.
- Sharing and circular economy-based co-creation will enable the promotion of sustainable interaction with existing resources and processes.
- Development of products and technologies that innovate to zero will be promoted, for example, zero-waste and zero-emission technologies.
- Immersive digital realities will facilitate novel ways of learning, understanding, and memorizing in several fields of science.

While three usage scenarios defined in IMT-2020, namely eMBB, mMTC, and URLLC, will continue to hold significance, there is a need to consider new use cases and applications to drive the ongoing evolution of technologies and meet future requirements. The emergence of novel technologies like holography, robotics, microelectronics, renewable energy, photo-electronics, AI, and space technology offers opportunities to introduce unprecedented products and services. In order to emphasize the distinctive features and outline the technical prerequisites, wireless researchers have endeavored to envision several potential killer applications for IMT-2030, which include,

Holographic-Type Communication (HTC) Holographic displays represent the next step in enhancing the multimedia experience by projecting 3D images from one or multiple sources to multiple destinations, providing an immersive and lifelike encounter for users. The network's interactive holographic capability demands extremely high data rates and exceptionally low latency. Unlike conventional 3D videos that rely on binocular parallax, authentic holograms replicate all visual cues of observing 3D objects with the naked eye, creating a natural and genuine 3D perception. Recent advancements in holographic display technology, exemplified by Microsoft's HoloLens, suggest that widespread application is likely in the next decade. Through a mobile network, remote rendering of high-definition holograms promises to deliver an incredibly immersive experience. For instance, holographic telepresence could project remote participants as holograms into meeting rooms, enabling realistic interactions during online training and education with ultra-realistic objects. However, the implementation of holographic technology leads to a substantial demand for bandwidth, potentially reaching terabits per second, even with image compression. Beyond typical aspects like frame rates, resolution, and color depth in two-dimensional (2D) video, the quality of hologram also involves the volumetric data such as tilt, angle, and position. If representing an object with images every 0.3°, an image-based hologram with 30° field of view and a tilt of 10° needs a 2D array of 3300 separate images (Clemm et al., 2020). Moreover, holographic technology also requires ultra-low latency to achieve true immersiveness and precise synchronization across numerous interconnected streams for accurate hologram reconstruction.

Tactile Internet Tactile Internet ensures an exceptionally low E2E latency, meeting the demanding 1 millisecond (ms) or even faster reaction time, which approaches the limits of human perception (Fettweis et al., 2014). This, combined with high reliability, availability, security, and at times, high throughput, leads to the emergence of disruptive real-time applications. These advancements play a crucial role in enabling real-time monitoring and remote control for Industry 4.0 and Smart Grid. Human operators can remotely monitor machines using VR or holographic devices while being assisted by tactile sensors that may involve actuation and kinesthetic feedback for better control. The typical closed-loop controlling, especially for devices or machinery rotating rapidly, is very time-sensitive, where an E2E latency below 1 ms is expected. In the healthcare domain, the tactile Internet facilitates numerous potential applications like tele-diagnosis, remote surgery, and telerehabilitation. Tele-

diagnostic tools and medical expertise become readily available anywhere and anytime, disregarding the patient and medical practitioner's physical locations. A prime example is remote and robotic surgery, where a surgeon receives real-time audio-visual feeds of the patient undergoing a procedure in a distant location.

Extended Reality Similar to video traffic that saturates the 4G networks, the proliferation of XR devices will be restricted by the limited capacity of 5G with the peak rate of 20 Gbps, especially at the cell edge. Extended reality, combining VR, AR, and MR, is expected to deliver enhanced features such as higher resolution, larger field of view, higher frames per second, and lower motion-to-photon latency. XR-enabled immersive applications such as gaming, remote surgery, and mission-critical remote manipulation require low latency and high reliability in addition to high data throughput. To achieve the same level of image quality, XR devices with 360° field of view need much higher data throughput in comparison to 2D video streaming. For an ideal immersion experience, the quality of video with higher resolution, higher frame rate, more color depth, and high dynamic range is required, leading to a bandwidth demand of over 1.6 Gbps per device (Huawei VR Report, 2018). One of the key challenges in supporting interactive XR experiences over cellular networks is the synchronized transport of multi-modal flows (e.g., visual media, audio, and haptics) to and from various devices in a collaborative group engaging with the same XR application. Additionally, facilitating real-time adaptations in the network according to user movements and actions is essential to ensure that interactions with other users and objects appear highly realistic in terms of placement and responsiveness. To enable spatial interactions, it is crucial to have fast accessibility and easy integration of content from different sources containing up-to-date and accurate representations of both real and virtual environments.

Digital Twin A digital twin is a virtual representation of a physical object, system, or process, meticulously crafted using real-time and historical data to mirror its real-world counterpart's behaviors, characteristics, and conditions. This cutting-edge technology has garnered significant attention in the context of the IoT, Industry 4.0, and advanced manufacturing, offering businesses and organizations unprecedented insights into their physical assets and processes. With the ability to simulate, analyze, and monitor these objects or systems in a virtual environment, digital twins are poised to revolutionize industries and drive innovation in ways previously unimaginable. While early deployments have already showcased remarkable potential, the full realization of the digital twin's capabilities is expected to coincide with the development of 6G networks, enabling even more sophisticated applications and transformations across diverse vertical industries and manufacturers. The digital twin's capacity to create detailed virtual replicas of physical entities, coupled with its potential for automation and intelligence, promises a future of enhanced efficiency and optimization across numerous domains. However, for its seamless integration and successful functioning, the demand for real-time, high-accuracy sensing, low latency, and high data transmission rates is essential to ensure real-time interaction between the virtual and physical worlds, enabling perceptive and cognitive intelligence and driving optimal outcomes through proactive management operations.

Pervasive Intelligence The widespread adoption of mobile smart devices and the emergence of connected technologies like robots, smart cars, drones, and VR glasses have paved the way for a surge in over-the-air intelligent services. These services heavily rely on traditional computation-intensive AI techniques such as computer vision, Simultaneous Localization And Mapping (SLAM), face and speech recognition, natural language processing, and motion control. However, due to the limited computing, storage, and connectivity resources of mobile devices, the full potential of these intelligent tasks can be realized with the advent of 6G networks. 6G networks will enable pervasive intelligence in an AI-as-a-Service approach (Letaief et al., 2019), leveraging distributed computing resources across the cloud, mobile edge, and end user devices while optimizing communication-efficient machine learning and interference mechanisms. For instance, humanoid robots like Boston Dynamics' Atlas can offload their SLAM computational load to edge computing resources, resulting in improved motion accuracy, prolonged battery life, and reduced weight by eliminating some embedded computing components. Beyond handling computation-intensive tasks, pervasive intelligence also enables time-sensitive AI tasks, bypassing the latency limitations of cloud computing when rapid decisions or responses are necessary.

Internet of Everything encompasses a wide array of scenarios, including real-time monitoring of diverse elements such as buildings, cities, the environment, transportation systems, roads, critical infrastructure, and water and power networks, among others. Additionally, the integration of the Internet of bio-things, represented by smart wearable devices and implanted sensors for intra-body communications, is expected to drive the demand for connectivity beyond mMTC. As we look to the future, there is a foreseen surge in the number of private networks, application-specific networks, and IoT networks. However, achieving seamless interoperability is one of the most significant challenges within this ubiquitous connectivity

and computing environment. With different products, processes, applications, use cases, and organizations interconnected, ensuring harmonious interactions among telecommunications networks, computers, and peripheral devices becomes crucial. The pursuit of effective interactions has been a subject of interest since the earliest days of distributed computing systems.

Multi-sense Experience While humans possess five senses – sight, hearing, touch, smell, and taste – to perceive the external environment, current communication technologies primarily focus on optical (text, image, and video) and acoustic (audio, voice, and music) media. However, the integration of taste and smell sensations into communication channels has the potential to create fully immersive experiences, opening new avenues for services in the gastronomy and texture industries. In addition, haptic communication, which involves the sense of touch, will play an increasingly crucial role in various applications such as remote surgery, remote control, and immersive gaming. Nevertheless, these use cases place strict demands on low E2E latency to ensure seamless and real-time interactions. Embracing the full spectrum of human senses in communication will undoubtedly drive innovation and open up exciting possibilities across a wide range of industries and experiences (ITU-T NET-2030, 2019).

Multi-dimensional Sensing Leveraging wireless signals for sensing, positioning, and imaging will unlock a realm of possibilities in various fields. This technology will enable high-precision positioning, ultra-high-resolution imaging, mapping, and environment reconstruction, as well as advanced gesture and motion recognition capabilities. These applications will require sensing systems with high resolution, exceptional accuracy, and efficient detection rates to deliver optimal performance and accuracy.

Intelligent Transportation and Logistics Automatic and connected vehicles represent a transformative technological shift in the transportation landscape. The integration of V2V and vehicle-to-infrastructure (V2I) communication and coordination, alongside autonomous transport capabilities, holds immense potential for reducing road accidents and traffic congestion. To achieve this, incredibly low latencies in the order of a few milliseconds will be crucial for enabling collision avoidance and supporting remote driving functionalities. Looking ahead to the year 2030 and beyond, we can envision a future where millions of autonomous vehicles and drones work in harmony to provide safe, efficient, and environmentally friendly transportation of people and goods. However, such a vision of connected autonomous vehicles also brings forth stringent demands on reliability and latency to ensure the safety of passengers and pedestrians alike. The successful deployment of unmanned aerial vehicles, particularly in swarms of drones, opens the door to a plethora of groundbreaking applications. As technology continues to advance, the convergence of these intelligent transport systems holds the promise of revolutionizing mobility, offering unprecedented benefits to society while necessitating robust and efficient network infrastructures to meet the complex demands of this transformative era.

Global Seamless Coverage By far, mobile communications primarily catered to densely populated urban areas and indoor environments, leaving a significant digital divide for a large population in remote, sparsely populated, and rural regions without access to essential information services. This disparity has created a clear gap among people across the globe. Additionally, with over 70 % of the Earth's surface covered by water, the growth of maritime applications demands network coverage for both the water surface and underwater. However, achieving ubiquitous coverage across the entire planet with sufficient capacity, acceptable QoS, and affordability remains a challenge. Terrestrial networks face technical limitations in reaching remote areas and extreme topographies like oceans, deserts, and high mountain regions. Moreover, providing communication services to sparsely populated areas becomes economically unfeasible due to high costs. Meanwhile, geostationary Earth orbit (GEO) satellites, although offering some capacity, are costly to deploy and mainly serve high end users such as maritime and aeronautic industries (Qu et al., 2017). To address these challenges, the deployment of a large-scale LEO satellite constellation has emerged as a promising solution, enabling low-cost and high-throughput global communication services. The envisioned 6G system aims to capitalize on the synergy of terrestrial networks, satellite constellations, and other aerial platforms, thereby achieving ubiquitous connectivity for global eMBB users and facilitating ubiquitous IoT applications. By harnessing the potential of non-terrestrial networks, the goal is to bridge the connectivity gap and create a truly interconnected world for all.

11.4 IMT-2030 Usage Scenarios

Starting from 1985, when IMT-2000, initially named FPLMTS, was developed, the ITU-R has been actively engaged in the standardization of terrestrial IMT technology for each generation. This has entailed the definition of visions for 4G (IMT-Advanced) and 5G (IMT-2020), in the form of recommendations:

- ITU-R M.1645 (2003) – *Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000*
- ITU-R M.2083 (2015) – *IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond*

In February 2013, the ITU-R WP 5D took the initiative to explore the IMT vision for 2020 and examine future technological trends for the terrestrial IMT systems in 2020. The goal of IMT-2020 was to address a wide range of QoS requirements that were not encountered by mobile subscribers in previous generations, as these were driven by vertical industries. The findings were later incorporated into the recommendation ITU-R M.2083 (2015), which defined three usage scenarios:

- **enhanced Mobile Broadband (eMBB):** Mobile broadband addresses the needs of human-centric use cases, providing access to multimedia content, cloud services, and data. As smart devices such as smartphones, tablets, and wearable electronics become more prevalent and the demand for video streaming increases, the importance of mobile broadband continues to rise. This trend led to the introduction of what ITU-R refers to as "enhanced mobile broadband." It encompasses two major scenarios, including wide-area coverage and hotspots, each with distinct requirements. In the case of hotspots, which involve areas with high user density, there is a significant need for very high traffic capacity. Mobility requirements are relatively low, while the user data rate is higher compared to wide-area coverage scenarios. On the other hand, wide-area coverage demands seamless connectivity and medium to high mobility, while the data rate requirement for wide-area coverage may be more relaxed compared to that of hotspots.
- **Ultra-Reliable Low-Latency Communications (URLLCs):** This scenario aims to provide support for mission-critical machine-type communications. It represents a significant advancement compared to previous generations of cellular systems, which primarily focused on human-centric services. This advancement opens up possibilities for delivering mission-critical wireless applications, including but not limited to automatic driving, vehicle-to-vehicle communication for safety purposes, wireless control of industrial manufacturing or production processes, remote medical surgery, distribution automation in smart grids, and transportation safety.
- **massive Machine-Type Communications (mMTCs):** This scenario enables massive connectivity by accommodating a large number of connected IoT devices that usually have infrequent transmissions of delay-tolerant data. These devices, such as remote sensors and monitoring equipment, are designed to be cost-effective and consume low power, ensuring an extended battery life of up to 10 years.

It is not hard to know that the three 5G usage scenarios do not fully support the disruptive use cases and applications of 6G. For instance, users employing lightweight VR glasses for interactive gaming demand not only ultra-high bandwidth but also low latency. Similarly, autonomous vehicles on roads or flying drones necessitate seamless and high speed connectivity with high reliability and low latency. Within the wireless community, discussions have been taking place to explore potential usage scenarios for 6G. There are a dozen of different definitions for possible usage scenarios. For example, Jiang et al. (2021) applied a holistic methodology to derive three novel scenarios, i.e., *ubiquitous Mobile Broadband (uMBB)*, *ultra-reliable low-latency broadband communication (ULBC)*, and *massive Ultra-reliable Low-Latency Communication (mULC)*. Huawei announced its vision toward beyond 5G system (Huawei NetX2025, 2021), which can improve real-time interaction experience for individual users, enhance cellular IoT capabilities, and explore new scenarios, including new scenarios like *Uplink-Centric Broadband Communication (UCBC)*, *Real-Time Broadband Communication (RTBC)*, and *Harmonized Communication and Sensing (HCS)*, for a better, intelligent world.

Likewise, the ITU-R started the development process for 6G by defining the vision of IMT-2030 as the first step. This initiative gained momentum in early 2021 when ITU-R WP 5D officially initiated the study for the new recommendation, known as ITU-R M. [IMT Framework for 2030 and Beyond], which was approved and formally named as ITU-R M.2160 Recommendation. A comprehensive full-day workshop titled "IMT for 2030 and beyond" was held on June 14, 2022, attracting a total of almost 400 participants. Throughout this event, ITU-R members, external organizations, research projects, and academic institutions presented 14 in-depth discussions, showcasing their profound interest and visions for IMT-2030. After 2 years and 4 months of continuous discussions and a total of 156 contributions from around the world, the draft of the

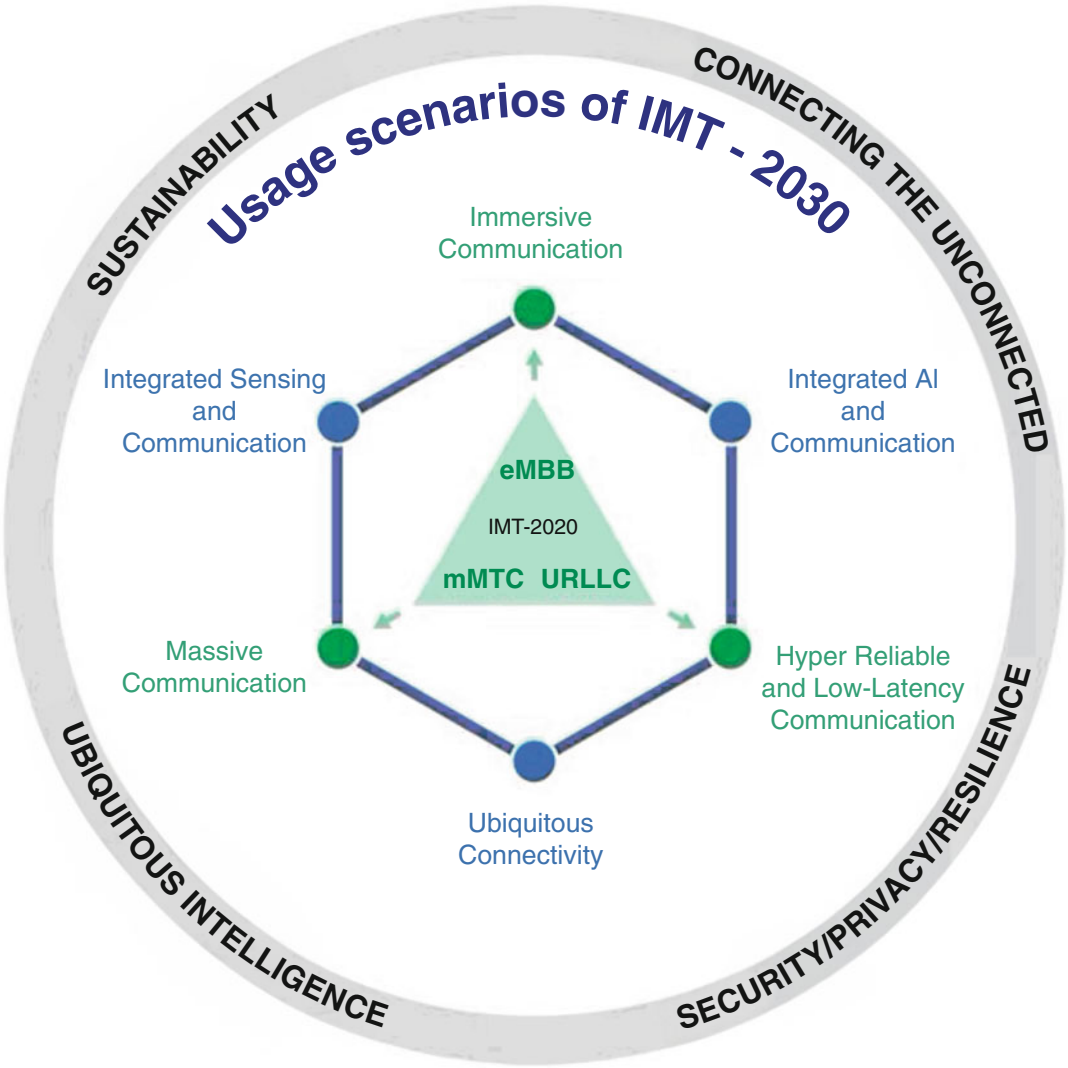


Fig. 11.2 Six usage scenarios for IMT-2030 and four overarching aspects. Source: ITU-R M. 2160

new recommendation for the IMT-2030 vision was completed on June 22, 2023 during the 44th ITU-R WP 5D meeting in Geneva.

Based on the potential use cases and applications, this framework recommendation specifies six usage scenarios of IMT-2030, as shown in Fig. 11.2, which extend from three 5G usage scenarios. On the circle around the hexagon are the four overarching aspects, i.e., sustainability, ubiquitous intelligence, security/privacy/resilience, and connecting the unconnected, that act as essential design principles for all usage scenarios.

- **Immersive Communication**: This usage scenario extends eMBB of IMT-2020, with a primary focus on delivering immersive experiences for users. It supports a wide range of use cases and applications such as XR, remote multisensory telepresence, and holographic communications. To achieve these immersive experiences, this scenario requires ultra-high transmission speed to enable real-time rendering of a large volume of visual information.
- **Hyper Reliable and Low-Latency Communication (HRLLC)**: This usage scenario extends the URLLC of IMT-2020 and covers specialized use cases that are expected to have more stringent requirements on reliability and latency.
- **Massive Communication**: This usage scenario extends mMTC of IMT-2020 and involves the connection of a massive number of devices or sensors for a wide range of IoT use cases and applications.
- **Ubiquitous Connectivity**: This particular usage scenario, one of the three newly introduced ones, primarily focuses on enhancing communication capabilities while placing significant importance on extending coverage and mobility. It

involves advancements in coverage extension technologies for terrestrial radio networking and will also introduce new architectures and business models to enable seamless interworking between terrestrial and non-terrestrial networks. By implementing this novel usage scenario, eMBB and IoT services will be expanded into rural, remote, and sparsely populated regions. This expansion is intended to be achieved at an affordable cost, ensuring that communication services become more inclusive and accessible in underserved areas. The ultimate goal is to connect all the currently unconnected people to bridge the digital divide.

- **Integrated AI and Communication**: This usage scenario would support distributed compute and AI-powered applications by leveraging data collection, local or distributed compute offload, and the distributed training and inference of AI models across various intelligent nodes. The power of AI made it possible to shift wireless communication to an intelligent paradigm, where architectures, protocols, and algorithms for the IMT-2030 will be designed by learning from wireless big data which has yet to be comprehensively exploited. In turn, with the wide deployment of base stations, edge servers, and intelligent devices, mobile networks will provide a novel and powerful platform for ubiquitous data collection, storage, exchange, and computing which are potential enablers for future mobile distributed/collaborative ML. For the future cellular system, an innovative and transformative shift will provide AI-as-a-Service to everyone, every business, every service anywhere anytime.
- **Integrated Sensing and Communication**: This usage scenario facilitates new applications and services that require sensing capabilities, which makes use of IMT-2030 to offer wide-area multi-dimensional sensing that provides spatial information about unconnected objects as well as connected devices and their movements and surroundings. The emergence of a sensing-as-a-network service is imminent, offering increased options to the market and a significant boost to various vertical industries. Use cases like real-time imaging of the surrounding environment, precise mapping, gesture and activity recognition, target detection and tracking, security surveillance, and navigation, as well as social welfare applications such as disaster monitoring, all stand to benefit greatly from the capabilities enabled by 6G signals.

11.5 IMT-2030 Capabilities

In order to well support disruptive use cases and applications in 2030 and beyond, IMT-2030 will be designed to offer high capacities and performance. Similar to the minimal performance requirements for IMT-2020, as specified in ITU-R M.2410 (2017), a set of quantitative and qualitative KPI will be defined to indicate the technical requirements for 6G. While most of the KPI used for evaluating 5G will remain applicable for 6G, there will be the introduction of some new KPI to assess novel technological aspects like sensing and intelligence. The 6G framework recommendation renames technical requirements or KPI in 5G with a new term called *Capabilities* and estimates the range of values for the capabilities. Figure 11.3 provides an overview of the various dimensions of capabilities for IMT-2030, encompassing nine enhanced capabilities (peak data rate, user-experienced data rate, spectrum efficiency, area traffic capacity, connection density, mobility, latency, reliability, and security/privacy/resilience) and six new capabilities (coverage, positioning, sensing-related capabilities, AI-related capabilities, sustainability, and interoperability). For each specific usage scenario, a single or multiple values within the range will be determined during the technical performance requirements phase, beginning in the year 2024.

- **Peak data rate** is the highest data rate under ideal conditions, e.g., all assignable radio resources (excluding radio resources for physical layer synchronization, reference signals, guard bands, and guard intervals) for the corresponding link direction are utilized for a single mobile station. It is proportional to the channel bandwidth and the peak spectral efficiency in that band. This requirement is defined for the purpose of evaluation in the eMBB usage scenarios. Traditionally, it is the most symbolic parameter to differentiate different generations of mobile systems. Driven by both user demand and technological advances, the peak rate of 6G is definitely greater than that of 5G, which has a minimal peak rate of 20 Gbps for downlink and 10 Gbps for uplink.
- **User-experienced data rate** is defined as the 5th percentile point (5%) of the cumulative distribution function (CDF) of user throughput. The user throughput is defined as the number of correctly received bits in the Service Data Units (SDUs) delivered to Layer 3, over a certain period of time during the active mode. In other words, a mobile user can get at least this data rate at any time or location with a possibility of 95%. It is more meaningful to measure the perceived performance, especially at the cell edge, and reflect the quality of network design such as site density, architecture, and inter-cell optimization. In the dense-urban deployment scenario of 5G, the target of user-perceived rate is defined as 100 Mbps in the downlink and 50 Mbps in the uplink. 6G is expected to offer much higher user-experienced rates than that of 5G, with a user-experienced rate like 300 Mbps or 500 Mbps.

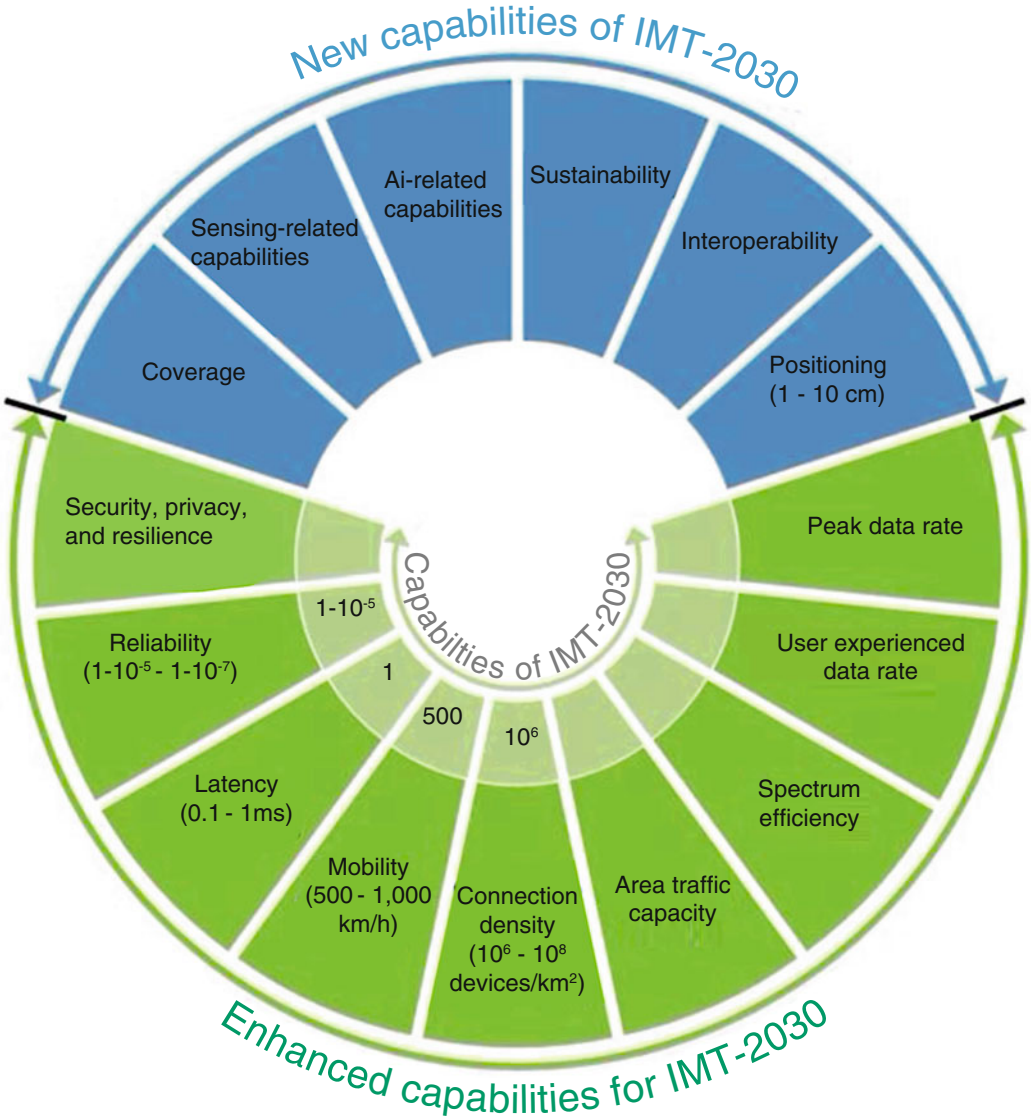


Fig. 11.3 Capacities for IMT-2030, consisting of nine enhanced capabilities and six new capabilities. Source: ITU-R M.2160

- **Spectrum efficiency** is an important capability to measure the advance of air interface technologies. The minimum requirement in IMT-2020 for peak spectral efficiencies is 30 bps/Hz in the downlink and 15 bps/Hz in the uplink assuming an antenna configuration to enable eight spatial streams in the downlink and four spatial streams in the uplink. Following the empirical data, it is expected that advanced 6G radio technologies with a higher spatial degree of freedom could achieve 1.5 or 3 times higher spectral efficiency over the 5G system.
- **Area traffic capacity** is a measurement of the total mobile traffic that a network can accommodate per unit area, relating to the available bandwidth, spectrum efficiency, and network densification. The minimal requirement for 5G is 10 Mbps per square meter (m^2), which is expected to improve several times, reaching 30 Mbps to 50 Mbps per square meter (m^2).
- **Connection density** is the capability applied for the purpose of evaluation in the usage scenario of massive communications. Given a limited number of radio resources, the minimal number of devices with a relaxed QoS per square kilometer (km^2) is 10^6 for the scenario of mMTC in IMT-2020. It is envisioned that 6G massive communication can support higher density than that of IMT-2020, falling in the range of 10^6 to 10^8 devices per km^2.

- **Mobility** means the maximal velocity of a mobile station supported by a network with the provisioning of acceptable QoS and quality of experience (QoE). To support the deployment scenario of high speed trains, the highest mobility supported by 5G is 500 km/h. Different classes of mobility are defined:
  - Stationery: 0 km/h
  - Pedestrian: 0 km/h to 10 km/h
  - Vehicular: 10 km/h to 120 km/h
  - High Speed Vehicular: 120 km/h to 500 km/h

  In IMT-2030, the mobility support will be further enhanced, with the maximal speeds of 500 km/h to 1000 km/h, which is targeted if commercial airline systems are considered.
- **Latency** can be differentiated into two categories: user plane latency and control plane latency. The former is the time delay induced in a radio network from a packet being sending out at the source until the destination receives it, assuming a mobile station is in the active state. To be specific, it is defined as the one-way time it takes to successfully deliver a small application layer packet (e.g., 0 byte payload with an IP header) from the radio protocol layer 2/3 SDU ingress point to the radio protocol layer 2/3 SDU egress point of the radio interface in either downlink or uplink in the network for a given service in unloaded conditions. In 5G, the minimum requirement for user plane latency is 4 ms for eMBB and 1 ms for URLLC. This value is envisioned to be further reduced, even to 0.1 ms. Control plane latency refers to the transition time from a most "battery efficient" state (e.g., the idle state) to the start of continuous data transfer (e.g., the active state). The minimum latency for the control plane should be 20 ms in 5G and is expected to be also remarkably improved in 6G.
- **Reliability** relates to the capability of transmitting a given amount of traffic within a predetermined time duration with high success probability. This requirement is defined for the purpose of evaluation in the usage scenario of URLLC. In 5G networks, the minimum requirement for the reliability is measured by a success probability of $1–10^{-5}$ when transmitting a data packet of 32 bytes within 1 ms given the channel quality of coverage edge for the deployment scenario of urban macro environment. It is expected to be improved up to two orders of magnitude, i.e., $1–10^{-7}$ or 99.999 99 % in the next-generation system.
- **Sustainability** is important to realize cost-efficient mobile networks and reduce the total carbon dioxide (CO_2) emission for green ICT, playing a critical role from the societal-economic respective. Energy efficiency of the network and the device can relate to the support for the following two features: (a) efficient data transmission in a loaded case and (b) low energy consumption when there is no data. After the early deployment of 5G networks, there are already some complaints about its high energy consumption although the energy efficiency per bit has been substantially improved in comparison with the previous generations. In 6G networks, sustainability would be a major focus so as to improve the energy efficiency per bit while reducing the overall power consumption of the mobile industry.
- **Positioning accuracy** of the 5G positioning service is better than 10 m. Higher accuracy of positioning has a strong demand in many vertical and industrial applications, especially in indoor environments that cannot be covered by satellite-based positioning systems. With the deployment of dual-functional (sensing and communication) radio stations and new transmission methods, the accuracy supported by 6G networks is expected to reach 10 cm or even 1 cm.
- **Coverage** in the definition of 5G requirement mainly focuses on the received quality of radio signal within a single base station. The coupling loss, which is defined as the total long-term channel loss over the link between a terminal and a base station and includes antenna gains, path loss, and shadowing, is utilized to measure the area served by a base station. In 6G networks, the connotation of coverage should be substantially extended considering that the coverage will be globally ubiquitous and will be shifted from only 2D in terrestrial networks to 3D in a terrestrial–satellite–aerial integrated system.
- **Security and privacy** are necessary for assessing whether the operation of a network is secure enough to protect infrastructure, devices, data, and assets. The main security tasks for mobile networks are *confidentiality* that prevents sensitive information from being exposing to unauthorized entities, *integrity* guaranteeing that information is not modified illegally, and *authentication* ensuring that the communicating parties are who they say they are. On the other hand, privacy becomes a high priority to address growing concern and privacy legislation such as the General Data Protection Regulation (GDPR) in Europe. Some capabilities can be applied to quantitatively measure security and privacy, e.g., the percentage of security threats that are identified by threat identification algorithms, with which the effectiveness of anomaly detection can be evaluated.

11.6 Development Roadmap Toward IMT-2030

In the last section of this chapter, we aim to provide the readers with an up-to-date advance of 6G explorations from representative institutions, i.e., the ITU-R and 3GPP, with the potential roadmap for definition, specification, standardization, and regulation, as demonstrated in Fig. 11.4. Following the remarkable success of ITU-R in advancing IMT-2000 (3G), IMT-Advanced (4G), and IMT-2020 (5G), a similar approach will be employed once more for the development of IMT-2030 (6G). After thorough deliberations spanning over a year, ITU-R has finalized the overall timeline for developing IMT-2030 in June 2022, consisting of three key stages:

- Stage 1 – Vision definition, scheduled for completion in June 2023, before the world radiocommunication conference in 2023 (WRC-23)
- Stage 2 – Requirements and evaluation methodology, targeted for completion in February 2026
- Stage 3 – Specifications, expected to be finalized by June 2030

At its meeting in February 2020, the ITU-R WP 5D decided to start the work plan for a report on future technology trends for the evolution of terrestrial IMT systems (ITU-R WP5D, 2020). It invited organizations within and external to the ITU-R to provide inputs for its June and October meetings in 2021. After around two and a half years, the ITU-R WP 5D published its outcomes in the report ITU-R M.2516 (2022) – "*Future technology trends of terrestrial International Mobile Telecommunications systems towards 2030 and beyond*." It elaborates on technology enablers for enhancing radio interface and radio access networks, as well as emerging technology trends for enabling new functions and services. In early 2021, the ITU-R WP 5D officially initiated the study of a new recommendation known as "*M. [IMT Framework for 2030 and Beyond]*," commonly referred to as the 6G vision. Through continuous discussions and the submission of 156 contributions, the ITU-R WP 5D successfully completed the report ITU-R M.2516 (2022) – "*Future technology trends of terrestrial International Mobile Telecommunications systems towards 2030 and beyond* " on June 22, 2023 during its 44th meeting held in Geneva. This recommendation serves as the initial and foundational milestone for the development of 6G, outlining the framework and overarching objectives of terrestrial IMT systems for the year 2030 and beyond.

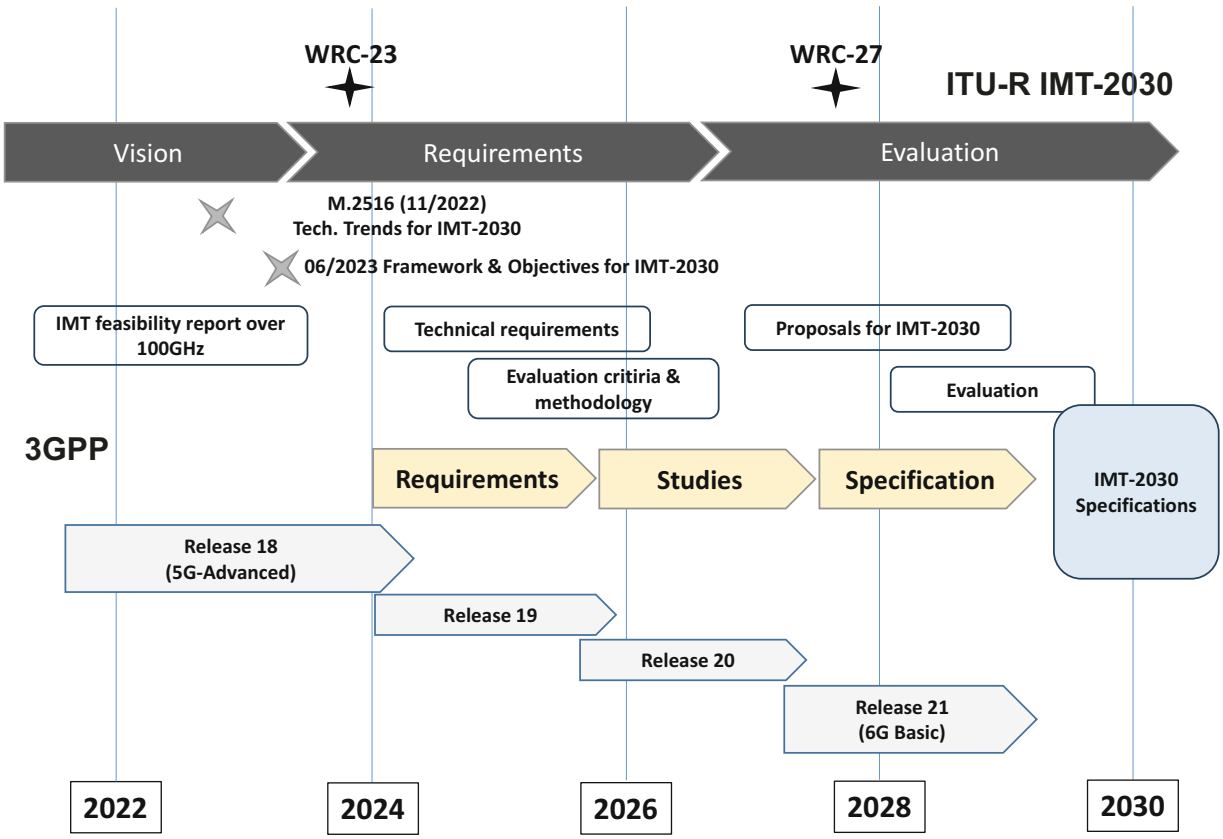


Fig. 11.4 Potential timelines of developing IMT-2030 in ITU-R and 3GPP

During the 41st meeting in June 2022, the ITU-R WP 5D reached an agreement on the timeline for finalizing the standardization of IMT-2030 radio interface technology by June 2030. Following the completion of the 6G vision, the working party is now shifting its focus toward defining detailed technical performance requirements, evaluation criteria, and methodologies of IMT-2030 over the next 3 years. Between 2027 and 2028, ITU-R WP 5D will accept submissions of candidate radio interface technologies from entities such as 3GPP and ITU member countries. These submissions will undergo rigorous evaluation to ensure compliance with the defined technical performance requirements. The selected IMT-2030 proposal(s) will be finalized and officially approved in June 2030. Additionally, ITU-R also bears the responsibility of organizing world radiocommunication conferences, which decide the frequency assignments and convenes every 3–4 years. To enable the commercialization of 6G by 2030, both technology development and standardization processes must be carried out in a timely manner. Simultaneously, securing the necessary spectrum for true 6G services requires proactive planning, with a focus on WRC-27, given its 4-year occurrence cycle. To determine the 6G spectrum in WRC-27, it is crucial to include candidate spectrum bands as agenda items during the forthcoming WRC-23, which is scheduled for the last quarter of 2023. The concerted efforts of all stakeholders will be required to secure the necessary 6G spectrum, and ITU-R remains committed to collaborating closely with them for this purpose.

In early 2019, the 3GPP has frozen the Release 15 specifications as the initial version of 5G standards. In July 2020, the subsequent release (i.e., Release 16) was completed as the enhancement of initial 5G standards (Ghosh et al., 2019). Release 17, the third version of 5G specifications, was frozen in mid-2022. Notably, this was the first and only release of 3GPP that was completed entirely remotely, using email discussions and online sessions, due to the travel restrictions imposed during the COVID-19 pandemic. 3GPP has introduced the official new name of *5G-Advanced* to signify the evolution of 5G in Release 18 and beyond (Chen et al., 2022). As of this writing, the specification of Release 18 is underway within 3GPP, with an anticipated completion date by the beginning of 2024 as per the current schedule. The specification of *5G-Advanced* is expected to be continued in the upcoming Release 19 and Release 20. According to the experiences got in previous generations, 6G should be a disruptive system that will be developed without the restriction of backward compatibility. 3GPP is expected to initiate the study item for 6G around the year 2025, followed by the phase of specification in Release 21, and prepare the technology proposal for IMT-2030 evaluation around late 2028 and early 2029, to guarantee the first commercial deployment roll-out of 6G by 2030.

11.7 Summary

Currently, both academia and industry in cellular communications have already initiated their intensified efforts to envision *what should 6G be* and discuss the potential key technologies. In June 2023, the first and fundamental milestone was achieved when the ITU-R successfully completed a new recommendation about the framework and overall objectives of International Mobile Telecommunications in 2030 and beyond (formally named IMT-2030), signifying a crucial step forward in the development of 6G. We aimed to shed light on the driving forces for developing 6G, predict explosive traffic growth by 2030, envision high-potential use cases and applications, report six IMT-2030 usage scenarios defined by ITU-R most recently, estimate performance capacities, summarize the research initiatives worldwide, and foresee the roadmap toward the deployment of 6G in 2030. Through studying this chapter, readers have an insightful view of the current state of the art and the road ahead in the evolution of cellular networks toward 6G and beyond.

11.8 Exercises

1. The new recommendation for the IMT-2030 vision was completed on June 22, 2023 during the 44th ITU-R WP 5D meeting in Geneva. This Framework recommendation specifies six usage scenarios of IMT-2030, compared with three 5G usage scenarios. Can you list these IMT-2030 usage scenarios?
2. Provide several research projects focusing on 6G.
3. Can you describe several key technologies that have a high potential for 6G?
4. Envision several future applications that need to be supported in 6G. Explain the reason that current 5G systems cannot accommodate these applications.

5. In addition to enhanced capabilities on the basis of IMT-2020 key performance indicators, such as peak data rate, mobility, area traffic capacity, and latency, IMT-2030 will introduce new capabilities reflecting the performance requirements on newly defined usage scenarios. Can you provide your observation about this connection between new capabilities and 6G usage scenarios?
6. In principle, three GEO communications satellites can cover the whole Earth's surface. Why *Ubiquitous Connectivity* is still a target for 6G?

Key Technologies for Sixth-Generation (6G) Mobile Cellular Systems

12

12.1 Technology Trends Toward IMT-2030

During the development process of 4G (IMT-Advanced) and 5G (IMT-2020), ITU-R studied the technology trends and released the results in 2004 and 2014, respectively, in the form of reports, i.e.,

- ITU-R M.2038 (2004) – *Technology trends*
- ITU-R M.2320 (2014) – *Future technology trends of terrestrial IMT systems*

Since the publication of ITU-R M.2320 in 2014, there have been significant advances in IMT technologies and the deployment of IMT systems. The capabilities of IMT systems are continuously updated in line with user trends and technological developments. Like the process in previous generations, the ITU-R WP 5D started a work plan for a report on future technology trends for the terrestrial IMT systems at its meeting in February 2020 (ITU-R WP5D, 2020). It invited organizations within and external to the ITU-R to provide inputs for its June and October meetings in 2021. After around two and a half years, the ITU-R WP 5D published its outcomes in the report ITU-R M.2516 (2022) – "*Future technology trends of terrestrial International Mobile Telecommunications systems towards 2030 and beyond*." It elaborates on technology enablers for enhancing radio interface and radio access networks, as well as emerging technology trends for enabling new functions and services.

In addition to a dozen of potential 6G technologies, ITU-R M.2516 (2022) also clarifies the drivers for future technology trends. According to this report, the evolution of the IMT systems and the underlying technologies must be guided by the imperative to satisfy fundamental requirements and contextualized in terms of how they can help society, the end users, and value creation. The non-technological driving factors include:

- *Societal goals* – Future technologies should contribute further to the success of several United Nations' sustainable development goals including environmental sustainability, efficient delivery of health care, reduction in poverty and inequality, improvements in public safety and privacy, support for aging populations, and managing expanding urbanization.
- *Market expectations* – New technologies should enable significant and novel capabilities, support radically new and differentiated services, and create greater market opportunities.
- *Operational necessities* – The need to manage complexity, drive efficiency, and reduce costs with end-to-end automation and visibility is also an imperative motivation and driving factor.

From a technological point of view, the development of IMT-2030 is mainly driven by the following aspects (ITU-R M.2516, 2022):

- *Energy efficiency* – This issue has long been an important design target for both the network and terminals. While improving energy efficiency, the total energy consumption should also be kept as low as possible for sustainable development. Power-efficient technology solutions are needed in both backhaul and local access to make use of small-scale renewable energy sources.
- *Data rate, latency, and jitter* – The data rate for future systems should be increased as much as practical in order to support extremely high bandwidth communications such as immersive XR and holographic services. Services with real-time and precise control usually have high demands on the latency of communications, such as the air interface delay, end-to-end latency, and roundtrip latency. Jitter refers to the degree of latency variation. Some of the future services such as time-

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5_12

sensitive industry automation applications may request zero jitter. Future systems should guarantee users' experience regardless of users' location and network traffic conditions.

- *Sensing resolution and accuracy* – Sensing-based services, including traditional positioning and new functions such as imaging and mapping, will be widely integrated with future smart services, including indoor and outdoor scenarios. Very high accuracy and resolution will be needed to support a better service experience.
- *Connection density* – It refers to the number of connected or accessible devices per unit space. It is an important indicator to measure the ability of mobile networks to support large-scale terminal devices. With the popularity of the IoT and the diversification of terminal accesses in specific applications such as industrial automation and personal health care, mobile systems need to have the ability to support ultra-large connections.
- *Coverage and full connectivity* – The forthcoming network is expected to offer worldwide coverage and seamless connectivity through a diverse multi-tier architecture. This comprehensive network should be capable of intelligent connectivity scheduling based on application needs and network conditions, enhancing resource efficiency and user experience. Additionally, it will expand the availability of reliable services like eMBB, massive IoT, and high-precision navigation across various environments, including urban and rural areas, and even extend into non-terrestrial spaces.
- *Mobility* – It refers to the maximum speed supported under a specific QoS requirement. Future systems will not only support terminals on land (e.g., high speed trains) but also provide services to terminals in airplanes, drones, and so on.
- *Spectrum utilization* – As we move toward 2030 and beyond, the introduction of new services and applications is likely to result in a significant increase in mobile data traffic. Consequently, there may be a need for additional spectrum to accommodate this explosive growth. To address this, further exploration into innovative applications of both low and mid bands, as well as the extension to higher frequency bands with broader channel bandwidth, should be pursued. Smartly utilizing multiple bands and enhancing spectrum efficiency through advanced technologies will be crucial in achieving high throughput within limited bandwidth constraints.
- *Simplified user-centric network* – In light of the vast array of new services and scenarios expected beyond 2030, the network must cater to a diverse range of demands and deliver personalized performance. To address this, the concept of *soft network* has been proposed, featuring a fully service-based and native cloud-based RAN. This soft network ensures QoS and maintains consistent user experience. The simplified user-centric network is a globally unified access network characterized by a straightforward architecture, boasting robust signaling control, precise network services, and efficient transmission. This is achieved through converged communication protocols and access technologies, allowing for plug-and-play and on-demand deployment. By adopting a user-centric approach, the network becomes fully distributed and decentralized, reducing the risk of a single point of failure. Furthermore, this user-centric model empowers users with control over their data ownership, a critical aspect of the next-generation network.
- *Native AI* – The upcoming mobile system is expected to possess enhanced capabilities and accommodate a broader range of services, leading to increased network complexity. AI reasoning will be seamlessly integrated into various aspects of the future network, including physical layer, radio resource management, network security, application enhancement, and overall network architecture. This integration will give rise to a multi-layer, deeply integrated intelligent network design. Moreover, the future network is anticipated to facilitate distributed AI-as-a-Service, enabling large-scale intelligence across the network infrastructure.
- *Security/trustworthiness* – In the future, networks are expected to provide more advanced system resilience to ensure reliable operation and service provision, robust security measures to guarantee confidentiality, integrity, and availability, and privacy features that enable self-sovereign control over data. Additionally, a focus on environmental safety will be vital. The roles of trust, security, and privacy are intertwined in various aspects of future networks. Both existing and emerging security threats must be effectively addressed. With the proliferation of novel IoT and other networked devices and their control systems, there will be ongoing concerns about security and privacy risks as well as new threat vectors. As we transition to IMT in the coming decades, it will be crucial to prioritize these aspects. Toward 2030 and beyond, IMT should aspire to support embedded end-to-end trust, significantly improving information security compared to the current network standards. This entails defining trust modeling, trust policies, and trust mechanisms.
- *Dynamically controllable radio environment* – A dynamically controllable radio environment may be able to change the characteristics of the radio propagation environment, therefore creating favorable channel conditions to support higher data rate communication and improve coverage.

12.2 Terahertz/THz Technologies

The abundance of spectrum available at THz frequencies presents a wealth of opportunities for ultra-fast wireless applications. It opens the possibility for terabit cellular hotspots, terabit campus/private networks, terabit D2D connectivity, holographic-type communications, untethered multi-modal XR goggles with viewport rendering at the network edge, and yet-to-be-conceived use cases. Moreover, it brings a new level of flexibility to mobile system design, as THz links can be harnessed for wireless backhaul/fronthaul among base stations, small cells, relays, remote radio heads, and roadside units. THz-based backhauling facilitates an ultra-dense network architecture, expedites network deployment, and reduces costs associated with site acquisition, installation, and maintenance. The tiny size of antennas used for transmitting THz signals, in the order of micrometers, holds the potential for wireless connectivity among nanoscale machines. Such nanomachines equipped with nanoscale antennas can serve specific tasks at the nanoscale, like a biosensor injected into a human blood vessel. Each component of these nanomachines typically ranges from a few hundred cubic nanometers to a few cubic micrometers for the entire device. It opens up exciting possibilities for unprecedented applications, including nanoscale communications among nanodevices, on-chip communications, the Internet of Nano-Things, and intra-body networks.

As we delve into higher frequencies, the spatial resolution of transmitted signals significantly improves, enabling precise spatial differentiation. In addition to THz communications, high-precision sensing and positioning take advantage of the incredibly small wavelengths on the order of micrometers and frequency-selective resonances of materials in the environment. These techniques extract unique information from the observed signal signature (Sarieddeen et al., 2020). Last but not least, another particular technical advantage is using THz radiation to form images. THz imaging exhibits high spatial resolution due to smaller wavelengths and ultra-wide bandwidths with moderately sized hardware than imaging using low frequencies. Compared with infrared lights, THz waves have better penetration performance, making common materials relatively transparent before THz imaging equipment. There are many security screening applications, such as checking postal packages for concealed objects, allowing THz imaging through envelopes, packages, parcels, and small bags to identify potential hazardous items. Based on the property that THz radiation is non-ionizing and therefore no known health risk to biological cells except for heating has motivated its application in the human body, where ionizing radiation, i.e., ultraviolet, X-Ray, and Gamma Ray, will raise high health risks.

The THz band presents extensive spectral resources for wireless communications, along with distinct advantages in sensing, positioning, imaging, and spectroscopy. This has led to a surge of interest in using it as a crucial component in implementing integrated sensing and communication (ISAC). These dual-functional wireless networks, which combine THz communications and THz sensing, offer a powerful synergy through two approaches: *sensing-aided communication* and *communication-aided sensing*. Leveraging sensing information in communications becomes a key advantage of ISAC, as it results in a more predictable and deterministic propagation channel. This, in turn, enables the design and implementation of efficient communications algorithms and protocols, such as sensing-aided channel estimation, predictive beamforming with sensing assistance, rapid beam alignment and tracking, and link blockage mitigation. On the other hand, mobile communications networks also present significant opportunities for network sensing or sensing as a service. By sharing sensing results through the mobile network, multiple network nodes like base stations and terminals can collaboratively act as a sensing system. This collaborative sensing, achieved through sensing data fusion, reduces measurement uncertainty and expands the coverage area, enhancing sensing accuracy and resolution.

Despite its high potential, the practical implementation of THz communications and sensing faces numerous challenges and issues. The most significant and immediate problem to address is the high propagation loss experienced in THz bands. Firstly, the receive antenna for the THz band has limited ability to capture radiation power due to the extremely small wavelength of THz waves. This leads to severe free-space path loss, which increases proportionally with the carrier frequency. Compounding this, the wavelength of THz waves is in the same order of magnitude as the dimensions of molecules in the atmosphere and human tissue. As a result, previously negligible phenomena like strong molecular absorption and particle scattering become significant concerns (Han et al., 2022). Specifically, water vapor and oxygen molecules in the atmosphere cause substantial losses, reaching up to approximately 20,000 dB per kilometer under worst-case conditions, as shown in Fig. 12.1. Additionally, suspended liquid water droplets in clouds, falling rain, hydrometeors, snowflakes, and fogs can also attenuate the signal strength due to their dimensions being comparable to the THz wavelength. Furthermore, surrounding physical objects become sufficiently large in size for scattering, and ordinary surfaces are also too rough to make specular reflections. As a result, a THz wave is susceptible to blockages like buildings, furniture, vehicles, foliage, and even humans.

The practical use of the THz band is challenged by large propagation losses, which generally lead to very short distances of signal transmission. This problem is further aggravated by the following two factors:

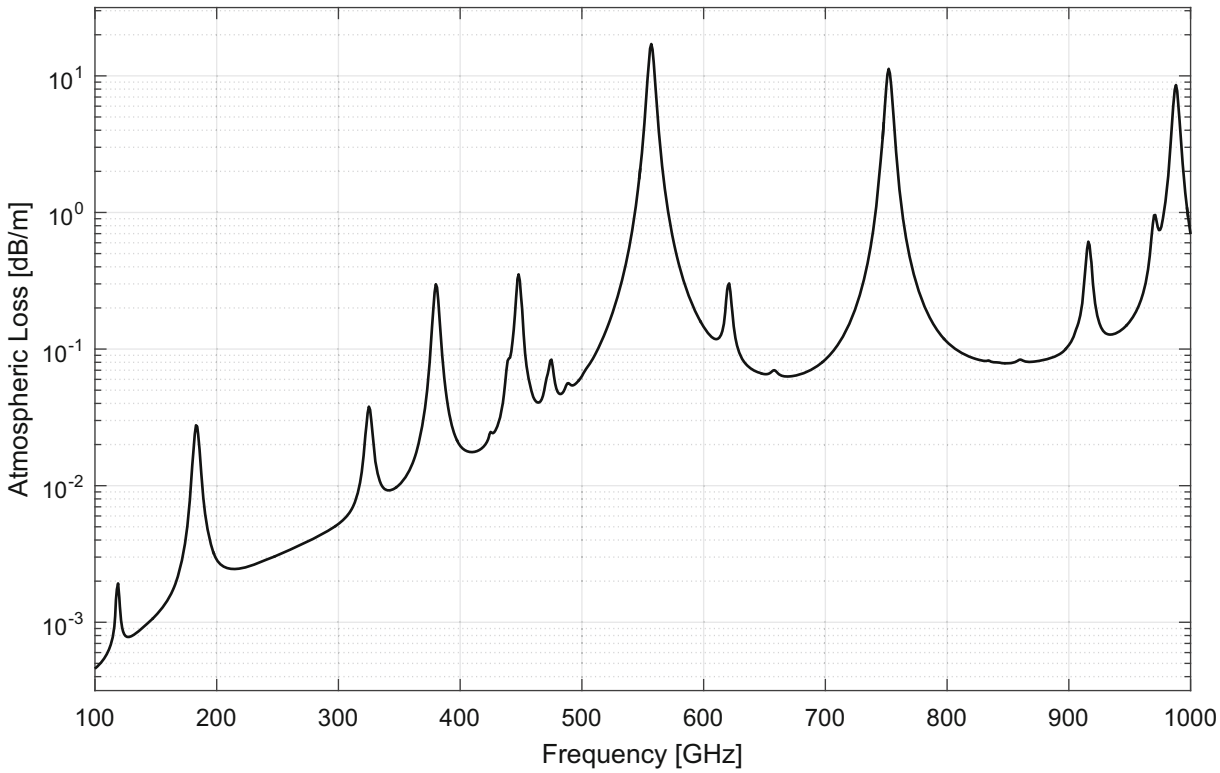


Fig. 12.1 Atmospheric attenuation from 100 GHz to 1 THz under the standard atmosphere condition with air pressure 1013.25hPa, temperature 15 °C, and water vapor density 7.5 g/m^3, according to ITU-R P.676 (2019)

- *Strong thermal noise*: Noise power is proportional to signal bandwidth with the constant power density. Therefore, the unique advantage of extremely wide bandwidth at the THz systems imposes a side effect of strong thermal noise.
- *Hardware constraint*: The transmit power at the THz band is quite constrained since the output power decreases with frequency and is at the level of decibel-milliwatts in the foreseeable future. Hence, raising power to extend the communications distance is not feasible.

To extend signal transmission distances beyond a few meters in THz communications and sensing, the use of high-gain directional antennas becomes essential to compensate for the high propagation loss. The advantage of tiny wavelengths allows for packing massive numbers of elements closely together in a small area, enabling high beamforming gains. The development of suitable pencil-beamforming algorithms plays a crucial role in THz-based wireless technologies, addressing challenges such as the huge number of antenna elements, beam squint, calibration requirements, and frequency windows. These algorithms focus on efficient device tracking in the THz band, leveraging accurate channel models and optimized signaling. THz technology requires highly directional antennas with beamforming capabilities, which may present challenges and losses in the antenna feed network. One alternative approach is the use of high-gain hemispherical lens antennas fed by a planar array with a moderate number of antenna elements. Other promising options expected to advance and reach practical implementation include graphene-based plasmonic antennas compatible at the nanoscale, plasmonic patch antennas, and graphene-based patch antenna arrays in muti-antenna configurations with beam-steering capabilities.

Micro- and macro-mobility are critical factors for the practical implementation of THz wireless links, particularly in mobile access scenarios. Even when users are not moving, slight rotation or movement of mobile devices can occur due to hand movements or other factors, necessitating efficient device tracking. Blockage of the line-of-sight link can also frequently happen, requiring device tracking to search for alternative paths. Beyond the physical layer, new link and network layer strategies are needed to deal with ultra-directional THz links. The requirement for highly directional antennas (or arrays) at both the transmitter and receiver introduces challenges in initial access strategies (Jiang & Schotten, 2022d,c), cell and user discovery, and relaying and collaborative networks. Receiver-initiated channel access policies based on polling from the receiver, instead of transmitter-led channel contention, are emerging as a potential solution. Additionally,

innovative strategies that exploit the full antenna radiation pattern for quicker neighbor discovery have been experimentally demonstrated. These aspects become even more challenging for specific use cases within THz communications and sensing.

12.3 Optical Wireless Communications/OWC

Optical wireless communications (OWCs) refer to a form of wireless communication that utilizes the infrared (IR), visible, or ultraviolet (UV) lightwaves as the transmission medium (Elgala et al., 2011). This technology shows great promise as a complementary solution to traditional wireless communications operating in RF bands. The use of the optical band allows for nearly limitless bandwidth without the need for approval from spectrum regulators worldwide. This, in turn, enables high speed and cost-effective access due to the widespread availability of optical emitters and detectors. One significant advantage of using IR and UV waves, which behave similarly to visible light, is the significant reduction of security risks and interference. Additionally, concerns regarding radio radiation's impact on human health are alleviated. As a result, OWC holds particular promise for deployment scenarios that are sensitive to electromagnetic interference, including intelligent transportation systems for vehicular communications, airplane passenger lighting, and medical equipment. Despite its advantages, OWC is not without its challenges. Impairments such as ambient light noise, atmospheric loss, nonlinearity of light-emitting diodes (LEDs), multipath dispersion, and pointing errors need to be addressed to ensure its optimal performance.

OWC systems operating in the visible light band within the frequency range of 400 THz to 800 THz are commonly referred to as visible light communications (VLC), attracted significant attention from both academia and industry in recent times. Unlike RF technologies that use antennas and operate in the lower THz range, VLC relies on LED and image sensor or photodiode arrays to function as transceivers. This approach allows VLC to achieve a large bandwidth with minimal power consumption (100 mW for data rates ranging from 10 Mbps to 100 Mbps), all while avoiding electromagnetic or radio interference. The advantages of VLC are numerous, including high power efficiency, long lifetime (up to 10 years), and cost-effectiveness of off-the-shelf LED. Additionally, VLC benefits from unlicensed spectrum access, making it an appealing solution for applications that prioritize battery life and access costs, such as massive IoT and wireless sensor networks. Furthermore, VLC demonstrates superior propagation performance compared to RF technologies in certain non-terrestrial scenarios like aerospace and underwater, which could prove vital for the future 6G ecosystem.

Terrestrial point-to-point OWC, also referred to as free-space optical (FSO) communications (Juarez et al., 2006), operates in the near-infrared band. By utilizing a high-power, tightly focused laser beam at the transmitter, the FSO system achieves a remarkable data rate of 10 Gbps per wavelength, even over long distances of up to 10,000 km. This technology provides a cost-effective solution to address the backhaul bottleneck in terrestrial networks. Additionally, it enables establishing connections among various platforms such as space, air, and ground, while also supporting high-capacity inter-satellite links, particularly beneficial for the emerging LEO satellite constellation. Moreover, there is a growing interest in UV communication, driven by recent advancements in solid-state optical transmitters and detectors. Non-line-of-sight UV communications offer broad coverage and enhanced security, contributing to the increasing enthusiasm for this technology.

In an OWC system, as illustrated in Fig.12.2, the transmitter modulates information bits into optical waveforms. Subsequently, the produced optical signals are transmitted through the atmosphere to a distant receiver. On the receiving end, a photodiode is employed to convert the received optical signal back into an electrical current. The transmitter consists of several components, including an optical source with its driving circuit, a channel encoder, a modulator, and a lens that focuses or shapes the optical beam. Initially, the information bits from the information source undergo encoding and modulation to form an electrical signal. This modulated signal then controls the optical intensity of the optical source. Prior to transmission, the optical beam is concentrated using an optical lens or beamforming optics. For short-range wireless communications, LEDs with beam collimators are commonly employed, whereas long-range transmission in FSO systems typically utilizes high-power laser diodes. These optical components must have a small footprint and low power consumption and deliver relatively high optical power over a wide temperature range with a long mean time between failures. Commercial FSO systems generally operate within two specific wavelength ranges: 850 nm and 1550 nm. These wavelengths coincide with the standard transmission windows of fiber optic communication systems and offer atmospheric attenuation of less than 0.2 dB per km. Consequently, off-the-shelf components are readily available for use in these wavelength ranges. Vertical-

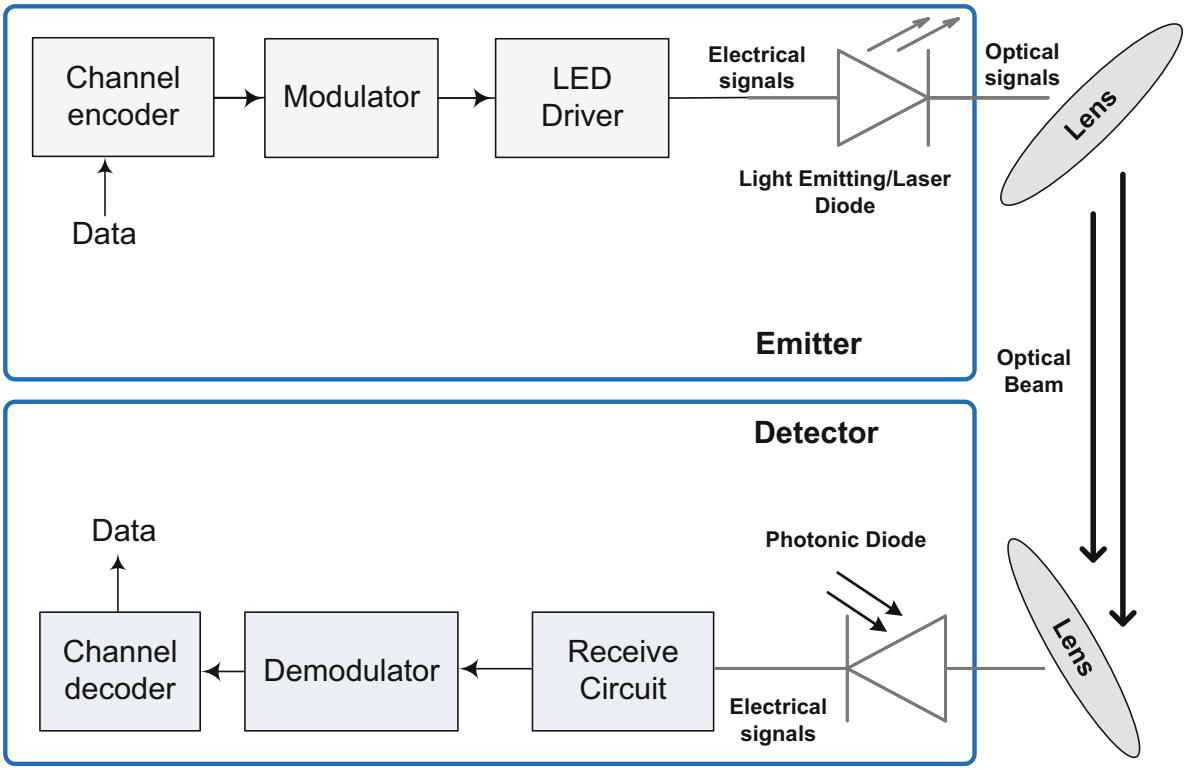


Fig. 12.2 Block diagram of an optical wireless communications system

cavity surface-emitting lasers (VCSELs) are primarily utilized for transmission around 850 nm, while Fabry–Perot (FP) and distributed feedback (DFB) lasers are predominantly used for operation at 1550 nm.

12.4 Ultra-massive MIMO

Given that the length of a resonant antenna is generally around the wavelength at the resonance frequency, the size of an array with tens of elements is several square meters at the sub-6G and mmWave bands and a few square centimeters at the THz band. As we move to the THz band, the antenna length decreases even further. It becomes possible to fit hundreds of elements compactly within a few centimeters using traditional metallic materials. Nevertheless, this quantity of elements is insufficient to counter the significant propagation loss experienced by THz signals (Akyildiz et al., 2018).

Utilizing surface plasmon polariton (SPP) waves allows for reducing the inter-element spacing of an array to the SPP wavelength, which is significantly smaller than the electromagnetic wavelength. This opens up the opportunity to use nanomaterials, such as graphene and metamaterials, which can support the propagation of SPP waves, to enhance hardware compactness. Graphene, an ultra-thin carbon nanomaterial with exceptional mechanical, electrical, and optical properties, is employed to manufacture plasmonic nano-antennas that are nearly two orders of magnitude smaller than traditional metallic THz antennas. At 1 THz, thousands of graphene-based nano-antennas can be integrated within a few square millimeters. The advent of nano-antennas has paved the way for creating large-scale arrays for THz communications. In 2016, Akyildiz and Jornet presented the concept of ultra-massive multi-input multi-output (UMMIMO) communications, demonstrating a remarkable MIMO system with dimensions of 1024×1024. Both the transmitter and the receiver were equipped with an array of 1024 nano-antennas, showcasing the potential of this technology (Akyildiz & Jornet, 2016).

Dealing with a large number of elements presents challenges like excessive power consumption and heightened hardware complexity in UMMIMO systems operating in the THz band. This situation calls for a reevaluation of the array architecture and beamforming strategies. Fully digital beamforming can create the desired beam but comes with the significant drawback of consuming excessive energy and incurring high hardware costs since each antenna in the large-scale array requires its own dedicated RF chain. It necessitates exploring analog beamforming as an alternative with lower complexity. By utilizing analog phase shifters, only one RF chain is necessary, resulting in substantial reductions in both hardware and

power costs. However, the analog architecture supports only a single stream, limiting data rates and the number of users. As a trade-off between these two approaches, a hybrid digital–analog architecture emerges as the optimal choice for THz systems. It combines an analog phase shifter network with a few RF chains, which significantly reduces hardware costs and energy consumption while achieving comparable performance to digital beamforming. Hybrid beamforming strikes a balance between performance and complexity, making it an attractive solution for THz signal transmission.

Although hybrid beamforming has been extensively studied for the sub-6GHz and mmWave bands, the peculiarities of the THz band, such as channel sparsity and beam squint, impose many difficulties for designing an UMMIMO system. Currently, many new forms of hybrid beamforming are discussed in the literature, including array of subarrays (AoSA) to balance the power consumption and data rate, widely spaced multi-subarray to overcome the low spatial multiplexing gain due to channel sparsity, and true-time-delay-based hybrid beamforming to address the problem of beam squint.

- *Array of Subarrays* – In a hybrid architecture, the connection between elements and RF chains has two basic forms: fully connected (FC) and AoSA. In the FC hybrid beamforming, each element is fully connected to all RF chains via a signal combiner, and the signal of an RF chain radiates over all antenna elements via an individual group of phase shifters. Any RF chain should have the capability to drive the entire large-scale antenna array, which is power-aggressive. Particularly, the use of a large number of phase shifters and combiners will exacerbate the problems of high hardware cost and power consumption. In contrast, all elements in AoSA are divided into disjoint subsets called subarrays, and a subarray is only accessible to one specific RF chain. AoSA conducts signal processing at a subarray level with fewer phase shifters, such that hardware cost, power consumption, and signal power loss can be dramatically reduced. In addition, beamforming and spatial multiplexing can be jointly optimized by cooperating with precoding in the baseband.
- *Widely Spaced Multi-subarray* – Due to the tiny wavelength, the THz channel is usually sparse, consisting of a LoS path and a few reflection paths. The transmit power concentrates on the LoS path, and the overall angular spread of THz signals is small. For instance, a maximal angular spread of 40° has been observed for indoor environments in the THz band, compared to 120° for indoor scenarios at 60 GHz mmWave frequencies. Since the number of spatial degrees of freedom is upper bounded by the number of multipath components, the number of data streams or the potential spatial multiplexing gain is much small, limiting the achievable data rate at the THz band. A widely spaced multi-subarray hybrid beamforming architecture is proposed by Yan et al. (2022) to overcome the low spatial multiplexing gain due to channel sparsity. Instead of critical spacing, the inter-subarray separation is over hundreds of wavelengths, reducing the correlation between the subarrays.
- *True-Time-Delay-Based Hybrid Beamforming* – Most of the current hybrid beamforming architectures rely on phase shifters, which are frequency independent, inducing the same phase rotation at different frequency components of a signal. Under the ultra-wide bandwidth at the THz, these shifters only provide correct phase shifting for a certain frequency point, whereas other frequency points suffer from phase misalignment. As a result, the formed beam is squinted with a substantial power loss. To solve the problem of beam squint at the THz band, true-time-delay (TTD) can be applied to substitute phase shifters. The TTD is frequency-dependent, and the phase rotation adjusted by TTD is proportional to the carrier frequency and perfectly matches the ultra-wideband THz beamforming.

12.5 Cell-Free Massive MIMO

Cellular networks hold promise with the adoption of an innovative wireless access technology called massive MIMO. This technology involves a base station equipped with a large-scale antenna array, enabling it to serve multiple users within a cell simultaneously, over the same time–frequency resource. The benefits of massive MIMO include enhanced throughput, reliability, and energy efficiency, achieved simply through linear signal processing techniques. In a massive MIMO setup, a large number of service antennas can be deployed in two ways: collocated or distributed configurations. Collocated architectures place all service antennas close together, resulting in lower backhaul requirements, simple synchronization, easy array calibration, and facilitating joint processing. However, it is essential to note that this configuration primarily benefits users located near the cell centers, leaving users at the cell edges with considerably inferior quality of service due to issues such as high signal attenuation, inter-cell interference, and handover challenges inherent to traditional cellular architecture. As cellular networks densify to accommodate higher system capacity, the issue of inter-cell interference and frequent handovers becomes more pronounced, further affecting overall network performance (Zhang et al., 2020).

As a result, the majority of traffic congestion in cellular networks occurs at the cell edge. The user-experienced performance, defined by the so-called 95%-likely user data rates, remains subpar in 5G networks, as it can only guarantee satisfactory performance to 95% of the users. To address these challenges, a potential solution is to establish multiple

connections between each user and a multitude of distributed antennas. This approach can be particularly effective when the network consists of a single large cell, eliminating inter-cell interference and the need for handovers. Various names have been used in the past to describe this solution, such as network MIMO, distributed MIMO, distributed antenna array, and CoMP transmission and reception. By efficiently leveraging spatial diversity to counteract shadow fading, a distributed antenna system can significantly enhance the probability of coverage compared to a collocated system. However, it is important to note that this improvement comes at the cost of increased backhaul requirements.

In the work by Ngo et al. (2017), they proposed a novel distributed massive MIMO system designed to serve a few users scattered across a wide geographical area using a large number of service antennas. These antennas work in unison through a fronthaul network, serving all users simultaneously over the same time–frequency resource. To minimize the overhead associated with acquiring CSI, the system operates in the TDD mode, exploiting channel reciprocity. This innovative system is distinctive in that it eliminates the notion of cells or cell boundaries, earning it the name *cell-free massive MIMO*. By combining the principles of distributed MIMO and massive MIMO, this setup is expected to harness the advantages of both systems effectively. Distributed MIMO enables improved spatial diversity and coverage, while massive MIMO enhances capacity and spectral efficiency, resulting in a promising wireless communication solution.

The typical setup of a cell-free massive MIMO system involves a large number of M access points (APs) and a smaller number of K users, where $M \gg K$. Basically, both the APs and user terminals are equipped with a single antenna and are randomly distributed across a given geographical area. These APs are connected to a central processing unit (CPU) through a backhaul network, as depicted in Fig.12.3. In this setup, all M APs simultaneously serve all K users using the same time–frequency resource. The downlink transmission occurs from the APs to the users, while the uplink transmission takes place from the users to the APs, with the TDD operation effectively separating the downlink and uplink. Each coherence interval within the system is divided into three phases: the uplink training phase, the downlink payload data transmission phase, and the uplink payload data transmission phase. During the uplink training phase, users transmit reference signals to the APs, and each AP independently estimates the channels with all users. By capitalizing on the channel reciprocity inherent in TDD systems, the base station acquires downlink channel knowledge from the estimated uplink CSI. These obtained channel estimates play a vital role in the system: They are utilized for precoding the transmitted signals in the downlink, as well as for detecting the signals transmitted from the users in the uplink.

We can write

$$g_{mk} = \sqrt{\beta_{mk}} h_{mk} \tag{12.1}$$

to model the fading channel between a generic AP $m = 1, \ldots, M$ and a typical UE $k = 1, \ldots, K$, where β_{mk} and h_{mk} represent large-scale and small-scale fading, respectively (Jiang & Schotten, 2021b). In a cell-free massive MIMO system, two linear precoding methods, i.e., conjugate beamforming and zero-forcing precoding, are applied to spatially multiplex the information-bearing symbols intended for K terminals, denoted by $\mathbf{u} = [u_1, u_2, \ldots, u_K]^T$, where the symbols are normalized

$$\mathbb{E}[|u_k|^2] = 1, \quad k = 1, 2, \ldots, K. \tag{12.2}$$

A cell-free massive MIMO system applying conjugate beamforming operates as follows:

- A typical AP m measures β_{mk}, $k = 1, 2, \ldots, K$, and reports them to the CPU. Usually, the large-scale fading keeps constant for a relatively long period with respect to the channel coherence time. Consequently, the knowledge of β_{mk} can be regarded as perfect, and the overhead of measurement and distribution is small.
- The CPU computes power control coefficients η_{mk}, $\forall m$, $\forall k$ as a function of β_{mk}, and sends them to corresponding APs. Meanwhile, the CPU distributes the information-bearing symbols $\mathbf{u}$ to all APs. Note that β_{mk} might be the same for tens of symbol periods, and therefore only $\mathbf{u}$ is needed to be transmitted.
- Users synchronously transmit their pilot sequences $\boldsymbol{\phi}_k$, $k = 1, \ldots, K$, with a duration of τ_p.
- The mth AP, where $m = 1, \ldots, M$, acquires the estimate of its own spatial signature $\hat{\mathbf{g}}_m = \left[\hat{g}_{m1}, \hat{g}_{m2}, \ldots, \hat{g}_{mK}\right]^T$.
- The APs treat the channel estimates as the true channels and use conjugate beamforming to generate the transmitted signals. The mth AP sends the signal

$$s_m = \sqrt{\eta_{mk} P_m} \hat{\mathbf{g}}_m^H \mathbf{u} = \sqrt{P_m} \sum_{k=1}^{K} \sqrt{\eta_{mk}} \hat{g}_{mk}^* u_k, \tag{12.3}$$

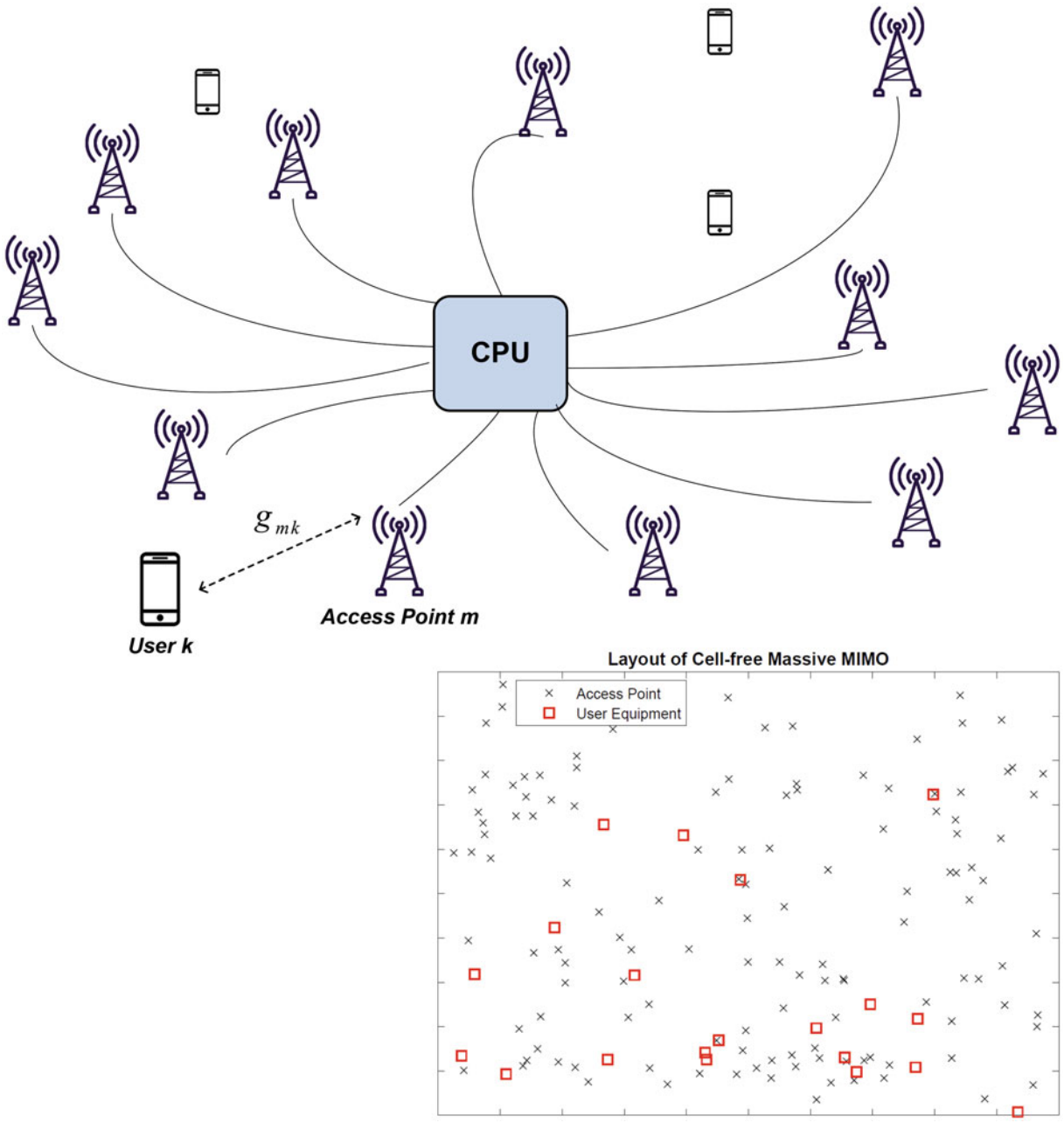


Fig. 12.3 A cell-free massive MIMO system where M single-antenna APs controlled by a CPU serves K terminals

where P_m is the transmit power limit of AP m. The choice of power control coefficients is subjected to

$$\mathbb{E}[|s_m|^2] \leqslant P_m, \quad \forall m = 1, 2, \ldots, M, \tag{12.4}$$

which can be interpreted as

$$\sum_{k=1}^{K} \eta_{mk} \mathbb{E}[|\hat{g}_{mk}|^2] \leqslant 1. \tag{12.5}$$

The philosophy behind the zero-forcing precoding is to completely suppress the interference among different users given the knowledge of downlink channels (Jiang & Schotten, 2021a). A cell-free massive MIMO system applying the zero-forcing precoding operates as follows:

- AP m measures β_{mk}, $k = 1, 2, \ldots, K$, and reports them to the CPU.
- As conjugate beamforming, the CPU computes power control coefficients in terms of β_{mk}. It is necessary to have $\eta_{1k} = \ldots = \eta_{Mk}$, $\forall k$, and therefore power coefficients should be only the functions of k, i.e., $\eta_{mk} = \eta_k$.
- Users synchronously transmit their pilot sequences $\boldsymbol{\phi}_k$, $k = 1, \ldots, K$.
- The mth AP, where $m = 1, \ldots, M$, acquires the estimate of its own spatial signature $\hat{\mathbf{g}}_m = \left[\hat{g}_{m1}, \hat{g}_{m2}, \ldots, \hat{g}_{mK}\right]^T$.
- Each AP sends its local CSI to the CPU, and therefore the CPU gets the global CSI $\hat{\mathbf{G}} = [\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2, \ldots, \hat{\mathbf{g}}_M] \in \mathbb{C}^{K \times M}$.
- The CPU jointly encodes the information-bearing symbols in terms of

$$\mathbf{s} = \hat{\mathbf{G}}^H \left(\hat{\mathbf{G}}\hat{\mathbf{G}}^H\right)^{-1} \mathbf{D}_\eta \mathbf{u}, \tag{12.6}$$

  where $\mathbf{D}_\eta \in \mathbb{C}^{K \times K}$ is a diagonal matrix consisting of power control coefficients, i.e., $\mathbf{D}_\eta = \text{diag}\{\sqrt{\eta_1}, \ldots, \sqrt{\eta_K}\}$.
- The CPU distributes the precoded symbol s_m to AP m, and these APs synchronously send their respective transmitted symbols toward the users.

12.6 Intelligent Reflecting Surface/IRS

Traditionally, stringent performance requirements in wireless communications, including ultra-high data rate, high energy efficiency, ubiquitous coverage, massive connectivity, ultra-high reliability, and low latency, as seen in 5G, have been achieved through three primary methods. The first involves deploying a dense and heterogeneous network with increased network equipment, which extends coverage and boosts capacity but leads to high costs, energy consumption, and interference. The second method integrates a massive number of antennas at base stations to harness spatial multiplexing gain, but it requires sophisticated signal processing, incurring high hardware costs and energy consumption. The third approach centers on wider signal bandwidths, reaching higher transmission rates but leading to spectrum shortage, and for future generations, utilizing higher frequency bands with challenges such as severe propagation loss and the need for denser networks and larger antenna arrays, exacerbating cost, energy, and interference issues. Therefore, there is a strong need to discover a disruptive and revolutionary technology that can sustainably enhance capacity and performance while remaining cost-effective, simple to implement, and energy-efficient.

On the other hand, a significant obstacle that severely constrains wireless communications' performance is the elusive nature of wireless channels, characterized by substantial path loss, shadowing, time variation, frequency selectivity, and multipath propagation. Traditional approaches to address this fundamental limitation involve compensating for channel loss and randomness through robust modulation, coding, and diversity techniques or adapting transmission parameters through adaptive control. Unfortunately, these techniques not only require substantial overhead but also have limited adaptability in the face of highly random wireless channels, creating a substantial barrier to achieving highly reliable wireless communications.

In this context, a promising and innovative solution known by various names such as intelligent reflecting surface (IRS) (Wu & Zhang, 2020), reconfigurable intelligent surface, large intelligent surface, large intelligent metasurface, programmable metasurface, reconfigurable metasurface, intelligent walls, and reconfigurable reflect-array has emerged to revolutionize the wireless channel propagation environment. Essentially, IRS is a planar surface comprising numerous small, passive, and cost-effective reflecting elements, as shown in Fig. 12.4. Each element has the capability to independently introduce a phase shift and/or amplitude attenuation (combinedly referred to as a reflection coefficient) to an incoming electromagnetic wave. Unlike conventional wireless transmission methods that adapt *passively* to the wireless propagation environment, IRS takes a *proactive* approach by intelligently controlling reflections, thereby enabling precise passive or reflective beamforming. By carefully designing its reflection coefficients, the signals reflected by IRS can be combined constructively with other signal paths to enhance the desired signal strength at the receiver or destructively to mitigate co-channel interference. This empowers wireless systems to have a new level of flexibility and adaptability with the assistance of a smart and programmable wireless environment.

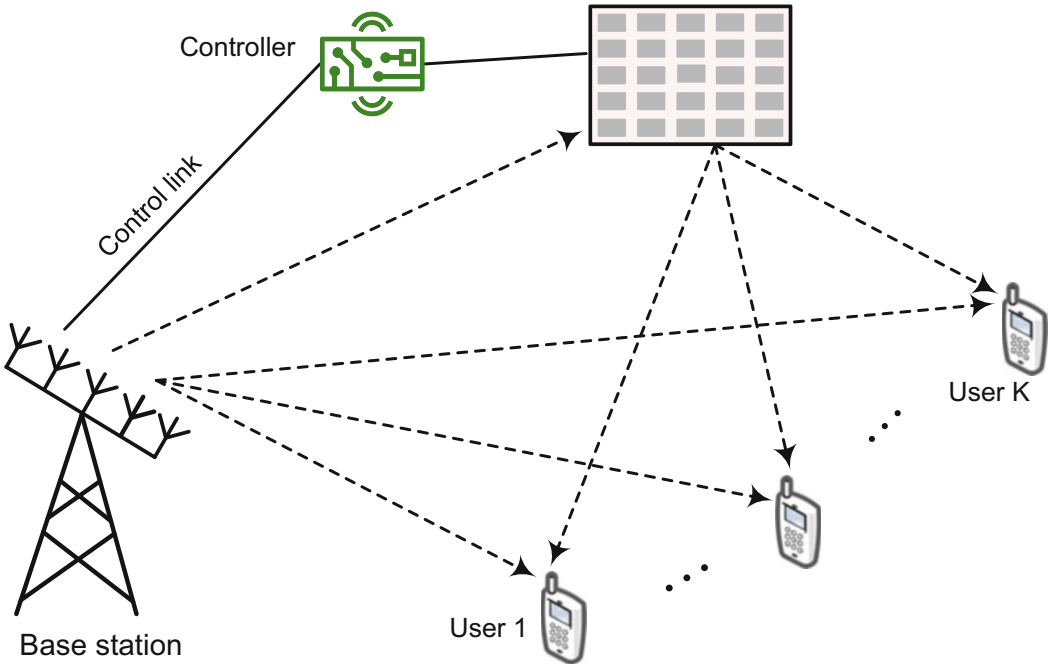


Fig. 12.4 Illustration of a typical IRS-aided wireless communication system

Due to its passive nature, where the reflecting elements (e.g., low-cost printed dipoles) only reflect electromagnetic waves without requiring RF chains for signal transmission and reception, IRS offers a significant advantage in terms of hardware cost and power consumption compared to traditional active antenna arrays. Implementing IRS can be accomplished with orders of magnitude less expense and energy usage. Additionally, the reflecting elements are typically lightweight and low profile and possess a conformal geometry. This allows IRS to be fabricated to adapt to various shapes, making it suitable for diverse deployment scenarios and easy integration into existing wireless networks as auxiliary equipment, ensuring excellent flexibility and compatibility. In summary, IRS is recognized as a disruptive technology characterized by its *low complexity, low cost, and low power consumption* while also holding the potential for high performance (Jiang & Schotten, 2022b).

IRS offers several distinct advantages when compared to other related technologies, namely wireless relays, backscatter communications, and active surface-based massive MIMO. Wireless relays typically operate in a half-duplex mode, leading to lower spectrum efficiency compared to IRS, which operates in a full duplex mode. While achieving a full duplex relay is possible, it requires sophisticated and expensive self-interference cancellation techniques. In contrast, IRS acts as a passive array that solely reflects electromagnetic waves without any active transmit module, such as a power amplifier, thereby avoiding noise amplification. Unlike traditional backscatter technologies like radio-frequency identification (RFID), which communicate by modulating their reflected signal emitted from the reader, IRS is primarily used to assist existing communication links without transmitting its own data. In backscatter communication, the reader must implement self-interference cancellation at its receiver to decode the RFID tag's message. In an IRS-aided transmission, both the direct link and the reflected link carry an identical signal, allowing them to be coherently superimposed at the receiver, thereby enhancing signal strength for improved detection. Furthermore, IRS differs from active surface-based massive MIMO in terms of their array architectures (passive vs. active) and operating mechanisms (reflecting vs. transmitting). These distinctive features position IRS as a highly advantageous and unique technology in the field of wireless communications (Jiang & Schotten, 2023b).

Given the aforementioned advantages, IRS offers great potential for massive deployment in wireless networks to significantly improve their spectral efficiency and energy efficiency while remaining cost-effective. This is expected to result in a fundamental shift in wireless system design, moving away from scaling up massive MIMO systems with a high number of antennas toward IRS-aided moderate-scale MIMO. Additionally, it will transition wireless networks from the existing heterogeneous setup to an IRS-aided hybrid network. In contrast to massive MIMO systems that rely on numerous active antennas to form sharp beams directly, an IRS-aided MIMO system enables a base station to operate with significantly fewer antennas while maintaining users' QoE through the use of smart passive reflection and exploiting the large aperture of IRS

to create precise reflect beams. This approach substantially reduces the system's hardware cost and energy consumption, particularly for wireless systems moving to higher frequency bands. Overall, IRS stands as a promising solution with the potential to revolutionize wireless networks and their efficiency on a large scale.

On the other hand, the current setup of wireless networks relies on a heterogeneous multi-tier architecture involving various active elements, such as macro, micro, and small base stations, remote radio heads, relays, and distributed antennas. These active nodes generate signals, which necessitate intricate interference coordination and cancellation mechanisms. Unfortunately, this conventional approach leads to increased network operation overhead, making it challenging to sustain the cost-effective growth of wireless network capacity. In contrast, by integrating IRS into the wireless network, a shift occurs from the existing heterogeneous network with solely active components to a novel hybrid architecture encompassing both active and passive components. IRS, being significantly more cost-efficient than their active counterparts, can be deployed more densely throughout the wireless network at a lower cost without causing interference, thanks to their passive reflection and localized coverage. By optimally balancing the ratio between active nodes and passive IRS in this hybrid network, a sustainable, environmentally friendly, and economically viable network capacity scaling can be achieved (Wu et al., 2021b).

12.7 Advanced Modulation and Coding

The capability of IMT-2030 is expected to significantly improve in terms of data rate, with an estimate of at least ten times compared to IMT-2020, reaching a target of up to 1 Tbps. However, there are challenges from the perspective of RF implementation, which is typically constrained by the limit of around 10% relative bandwidth. Even in the upper mmWave region (i.e., 100 GHz to 300 GHz), a single RF transceiver can only support a maximal bandwidth of 30 GHz due to the nonlinearity of RF components. To achieve the desired data rate of 1 Tbps, a bandwidth of 1 THz would be required with binary modulation. In contrast, higher-order modulation can reduce the necessary bandwidth, with 64-QAM requiring around 170 GHz. Nevertheless, this still presents challenges as the highest RF bands considered for IMT-2020 usage are below 100 GHz, and commercial deployment of lower mmWave bands starting from around 24 GHz is still ongoing. One proposed solution to meet the 1 Tbps communication requirement is to divide the necessary bandwidth into multiple parallel, non-interfering orthogonal sub-channels. This approach would involve splitting the bandwidth into at least six, and ideally more, separate sub-channels. To make these advancements possible, research into new modulation methods and signal-shaping schemes has become essential (ITU-R M.2516, 2022).

In addition to novel modulation and waveform design, new channel coding schemes also need to be studied. Emerging applications demand innovative channel coding schemes to meet different, sometimes conflicting, requirements: faster data rates, higher reliability, lower latency, massive connection density, reduced complexity, and lower power consumption. Given these diverse demands, researchers need to investigate advanced channel coding schemes, including enhanced versions of Polar codes, LDPC codes, and other coding technologies. These advanced codes should exhibit superior performance across a wide range of code lengths and rates while supporting flexible decoder choices, preferably unified into a single framework. Achieving higher throughput than legacy IMT systems requires considering both code design and corresponding algorithms to reduce decoding complexity and enhance decoding parallelism. Furthermore, new channel coding must maintain reasonable power consumption levels. Considering the significantly increased throughput requirements, the energy consumption per bit needs to be reduced further by one to two orders of magnitude. These new coding strategies should incorporate innovative iterative re-transmission mechanisms, especially for applications relying on short packets, such as those in IoT systems. While LDPC and Polar codes with short block lengths have been employed for IMT-2020 in traffic and control channels, respectively, using such codes poses challenges. On the one hand, short block length codes are less reliable, making error-free transmission difficult to guarantee. Increased error probability may necessitate re-transmissions, which may not be suitable for time-sensitive applications requiring low latencies. On the other hand, longer block length codes directly introduce higher latency. Hence, optimizing the interplay between the minimum required block length and robustness against transmission errors becomes crucial.

12.8 Next-Generation Multiple Access/NGMA

Multiple access technology plays a crucial role in facilitating numerous users to share radio resources effectively, serving as a cornerstone for the evolution of cellular standards. Multiple access can be categorized into two types based on resource sharing: orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA). Furthermore, it can also be classified based on the access procedure as grant-based multiple access and grant-free multiple access.

Both IMT-Advanced and IMT-2020 utilize OFDMA as their downlink multiple access technique. When compared to CDMA used in many IMT-2020 systems, OFDMA demonstrates significant advantages in mitigating multipath fading through simple one-tap channel equalization (Jiang & Kaiser, 2016b). Moreover, when combined with MIMO, OFDMA outperforms CDMA by a large margin in terms of spectral efficiency, flexibility, and scalability. As MIMO dimensions increase, transitioning from LTE-Advanced's maximum of 8×4 to over 256×32 for massive MIMO, and eventually moving toward future ultra-massive MIMO configurations like 1024×64, the complexity of MIMO-OFDMA systems grows exponentially and reaches levels that become impractical to handle. In comparison to OMA used in most conventional cellular networks, NOMA has the potential to provide higher system throughput and accommodate increased connection densities. NOMA achieves this by allowing multiple users to share the same radio resource unit. One specific instance of NOMA, known as multi-user superposed transmission (MUST), was extensively studied in LTE-Advanced Release 13, with a primary focus on downlink transmission. Additionally, Release 14 LTE introduced grant-based NOMA to support downlink eMBB. In Release 16, the exploration of NOMA shifted to uplink grant-free transmission, aiming to reduce control signaling overhead, transmission latency, and devices' power consumption.

Grant-based multiple access involves the coordination of users by a central unit before transmissions, assigning each user a unique signaling for receiver detection. While it requires dedicated protocols for communication coordination among accessible users, this technology, designed for conventional human-centric wireless networks, may not be suitable for thing-centric wireless networks supporting massive connection in cellular networks. Grant-free multiple access eliminates the need for extensive user coordination, making it more efficient in handling low latency, scheduling information deficiencies, and bursty user activity patterns. Though primarily used for initial access, grant-free multiple access faces challenges, including accommodating massive bursty devices, developing low complexity and energy-efficient coding and modulation schemes, and creating efficient detection methods for sporadic transmissions from a small number of active users.

In order to satisfy the rigorous and varied demands of IMT-2030, the development of advanced multiple access schemes, recently named next-generation multiple access (NGMA), is anticipated. NGMA is expected to efficiently support a vast number of users while conserving resources and minimizing complexity compared to existing multiple access methods. Based on the current state of the art in this research field, as presented by Liu et al. (2022), some potential candidates for NGMA include:

- *Power-domain NOMA (PD-NOMA)* – The core concept of PD-NOMA involves serving multiple users using the same time, frequency, or code while distinguishing them in the power domain. PD-NOMA relies on two key technologies: superposing coding and SIC. To address broadband communications over frequency-selective fading channels, where the channel coherence bandwidth is smaller than the system bandwidth, PD-NOMA can be seamlessly integrated with OFDMA. This integration involves assigning multiple users to each OFDMA subcarrier and serving them using PD-NOMA.
- *Code-domain NOMA (CD-NOMA)* – Taking inspiration from CDMA, which serves multiple users through shared time/frequency resources using dedicated user-specific spreading sequences, CD-NOMA was introduced. The main concept behind CD-NOMA remains the same - serving multiple users at the same time–frequency resources. However, CD-NOMA utilizes user-specific spreading sequences, which can be either sparse sequences or non-orthogonal cross-correlation sequences with low cross-correlation. At the receiver, multi-user detection is typically performed iteratively using Maximum Posteriori based algorithms.
- *Space-division multiple access (SDMA)* – By deploying multiple antennas at transmitters and/or receivers, it becomes possible to utilize additional spatial domain compared to communication systems with single antennas. This enhancement allows for serving multiple users in the same time/frequency/code domain while distinguishing them in the spatial domain. Linear precoding is the most commonly employed method for SDMA due to its computational simplicity. Specifically, linear precoding effectively mitigates inter-user interference by leveraging the spatial domain to design appropriate transmit and/or receive beamformers.
- *Rate-splitting multiple access (RSMA)* – RSMA is an innovative multi-antenna transmission scheme, introduced in recent years to facilitate multi-user multi-antenna communications. This approach makes use of the rate-splitting technique. In RSMA, the transmitter divides a portion of each user's message, known as the private message, into a common message that is intended for all users in the system. The remaining private messages, along with the newly formed common

message, are transmitted using beamformers, similar to the concept of SDMA. Upon reception, the receivers treat the private messages from all users as interference while decoding the common message. They then subtract the common message from the received signal using SIC. Subsequently, the intended private message is decoded and combined with a portion of the decoded common message to retrieve the complete information.

12.9 Open Radio Access Networks/O-RAN

In order to accommodate disruptive use cases, the next-generation network infrastructure needs to possess flexibility, intelligence, and openness for multi-vendor equipment and multi-tenancy. To achieve this goal, concepts such as softwarization, cloudification, virtualization, and network slicing will be fine-tuned for integration into the 6G network. Recently, a new network paradigm known as open radio access network (O-RAN) has gained much attention from both the academic and industrial sectors. The essential concepts of O-RAN, encompassing its vision, overall objectives, architecture, interfaces, enabling technologies, and other pivotal aspects, were initially introduced through the first white paper of the O-RAN alliance published in October 2018, entitled "*O-RAN: Towards an Open and Smart RAN*" (Chih-Lin et al., 2018). Subsequently, the O-RAN alliance delved deeper into exploring use cases that leverage the O-RAN architecture to showcase its real-time capabilities. The main objective of the openness and intelligence in RAN architecture is to build a radio network that is resource-efficient, cost-effective, software-driven, virtualized, slicing-aware, centralized, open-source, open hardware, intelligent, and therefore more flexible and dynamic than any previous generation of mobile networks. To achieve this ambition, the research community has introduced the incorporation of AI and machine learning techniques across all layers of the RAN architecture. This adaptation aims to fulfill the demands of densely populated network edges in the realm of 6G cellular systems.

Enabling the transition of the RAN from a closed environment dominated by a single vendor to a standardized, open, multi-vendor structure with AI-driven hierarchical organization grants the opportunity for third-party entities and mobile operators to implement innovative applications and emerging services that were unfeasible within traditional RAN setups. Additionally, the O-RAN framework is constructed upon the NFV management and orchestration or NFV-MANO reference architecture presented by ETSI. This architecture utilizes off-the-shelf commercial hardware components, virtualization methodologies, and software modules. The abstraction of virtual machines from the underlying physical resources facilitates effortless creation, deployment, configuration, and decommissioning processes. Consequently, this virtualized environment contributes flexibility to the O-RAN design, resulting in reduced CAPEX, OPEX, and energy usage. In summary, O-RAN is able to bring several advantages and benefits to the telecommunications industry, including:

- *Vendor Neutrality and Diversity*: O-RAN promotes an open ecosystem where different vendors can contribute components and solutions. This reduces vendor lock-in and encourages competition, leading to more innovation and diversity in network equipment and services.
- *Interoperability*: O-RAN defines open interfaces and standards that enable different components from various vendors to work seamlessly together. This interoperability simplifies network integration, reduces deployment complexities, and fosters a more flexible network architecture.
- *Cost Efficiency*: By allowing operators to mix and match components from different vendors, O-RAN can potentially lead to cost savings. Operators can choose the best-in-class solutions for different parts of the network, optimizing their investments.
- *Faster Innovation*: The open architecture of O-RAN allows for quicker development and deployment of new technologies and services. Operators can adopt innovations more rapidly without waiting for a single vendor's proprietary solutions.
- *Network Flexibility and Scalability*: O-RAN's modular approach enables network operators to adapt and scale their networks more easily. As technology evolves, operators can upgrade individual components without overhauling the entire network.
- *Virtualization and Cloud-Native*: O-RAN embraces virtualization and cloud-native principles, allowing network functions to be implemented as software running on standard hardware. This leads to better resource utilization, efficient scaling, and dynamic allocation of network resources.
- *Reduced Time-to-Market*: With standardized interfaces and open specifications, network operators can accelerate the deployment of new services and features. This is particularly important in the fast-paced world of telecommunications.
- *Easier Network Management and Maintenance*: O-RAN's open interfaces enable better visibility into network operations, making it easier to manage, monitor, and troubleshoot network components.

- *Community Collaboration*: O-RAN is developed through collaboration among network operators, equipment vendors, and other stakeholders. This collective effort fosters shared knowledge and best practices, leading to more robust and reliable networks.
- *Support for Rural and Underserved Areas*: O-RAN's flexibility and cost-effectiveness can make it more feasible to deploy and maintain networks in remote or economically challenged areas, helping bridge the digital divide.
- *Regulatory and Policy Compliance*: O-RAN's open architecture aligns with certain regulatory requirements and policies that encourage fair competition, data privacy, and security.

Despite the advantages in flexibility, interoperability, openness, and intelligence that O-RAN brings, it is not without its challenges, necessitating further research endeavors to fully realize its potential in upcoming mobile networks. These challenges encompass aspects such as aligning multi-vendor technologies on a common platform, harmonizing diverse management and orchestration frameworks, and effectively addressing performance concerns within the network. Both industry and academia experts are expected to collaborate in conducting theoretical analyses and practical implementations of this technology. These collaborative efforts aim to advance the development of an open and intelligent RAN for 6G mobile networks, ultimately surmounting these obstacles.

12.10 Non-Terrestrial Networks/NTN

The primary emphasis of traditional cellular systems was directed toward land-based infrastructure, resulting in a challenge pertaining to remote area coverage. In regions such as marine environments, oceanic expanses, and remote areas where establishing coverage through terrestrial cellular networks is either unattainable or economically unfeasible, satellites have long served as the predominant communication solution. In addition to this coverage problem, traditional terrestrial networks also face challenges in specific environments like emergencies and natural disasters. To address this need, the integration of non-terrestrial infrastructure into the 6G network has emerged as a significant topic. The new-generation networks envision a composition of three distinctive layers:

- The ground layer, constructed through terrestrial base stations
- The airborne layer, enhanced by high altitude platforms (HAPs) and UAV
- The spaceborne layer, established via a constellation of communication satellites

Terrestrial networks, covering a limited portion of the Earth's surface, encounter several limitations. Firstly, establishing terrestrial base stations to provide extensive coverage in areas such as oceans and deserts is technically unfeasible. Secondly, conquering challenging terrains like high mountains, valleys, and cliffs poses difficulties while using a terrestrial network for sparsely-populated regions is not cost-effective. Thirdly, terrestrial networks are susceptible to natural disasters like earthquakes, floods, hurricanes, and tsunamis. These situations necessitate communication, but the infrastructure may be destroyed or rendered non-operational. The expansion of human activities, such as passengers on commercial planes and cruise ships, has driven a growing demand for mobile Internet services in underserved regions. Additionally, the connectivity requirements of remote IoT deployments – such as environmental monitoring in mountain areas, offshore wind farms, and smart grids – call for ubiquitous coverage. While satellite communications have already provided a solution for wide coverage, where a single GEO satellite can cover around one-third of the Earth's surface, the current mobile communication services delivered by GEO satellites suffer from high costs, low data rates, and significant latency.

In contrast, LEO satellites offer distinct advantages over GEO satellites in delivering communication services, as highlighted by previous research (Hu & Li, 2001). Operating at altitudes generally below 1000 km, LEO satellites significantly reduce latency caused by signal propagation when compared to GEO satellites situated at an orbit of around 36 000 km. Additionally, the propagation loss in LEO orbits is notably lower, allowing for direct connectivity to battery-constrained mobile and IoT devices. Notably, stationary ground terminals like IoT devices positioned for monitoring might encounter line-of-sight obstacles from GEO satellites. An initial endeavor to establish a global satellite mobile communication system – the Iridium constellation – became commercially available in November 1998. Comprising 66 LEO satellites positioned at an altitude of around 781 km, this constellation provides mobile phone and data services across the entire Earth's surface. Despite facing challenges such as high costs and insufficient demand, the Iridium system remains a significant technological advancement and continues to operate today, with its second-generation system successfully deployed in 2019.

Recent years have witnessed the rise of SpaceX, a high-tech company gaining substantial attention for its revolutionary advancements in space launch technologies. Their reusable rockets, specifically Falcon 9, have substantially lowered the costs

associated with space launches, thereby creating opportunities for deploying large-scale space infrastructures. The Starlink system aims to finally deploy a large-scale LEO communications satellite constellation to provide ubiquitous Internet access services worldwide. Since 2015, it has already deployed over 4,000 Starlink satellites to serve a large number of users. The FCC has approved the initial plan to launch 12,000 satellites, and an application for an additional 30,000 satellites is currently being considered. In envisioning a future characterized by ubiquitous global connectivity available anytime and anywhere, it is strongly recommended to incorporate satellite networks as an integral component of the 6G network. This integration holds the potential to further advance the realization of such a converged global communication landscape.

12.11 Integrated Sensing and Communications/ISAC

Wireless sensing, encompassing tasks such as object detection, ranging, positioning, tracking, and imaging, has historically developed as a parallel technology track alongside IMT systems. IMT-2020 currently offers only positioning as a sensing service. However, there is a departure from the traditional approach of designing wireless networks solely for communication. Looking ahead to 2030 and beyond, IMT is envisaging the inception of an integrated sensing and communication system. This transformative shift will integrate sensing and communication functions right from the outset, resulting in a dual-functional network. In forthcoming communication systems, driven by novel features like the potential utilization of very high-frequency bands (ranging from mmWave to THz), wider bandwidth, denser network deployment, larger antenna arrays, and the integration of AI and collaboration among communication nodes, wireless sensing will emerge as a fresh function integrated seamlessly into the communication system. This integration will facilitate innovative services and solutions with a heightened degree of precision (Sarieddeen et al., 2020).

Within the ISAC paradigm, the mutual benefits of sensing and communication functions are underscored within a unified system, offering a powerful synergy through two approaches: *sensing-aided communication* and *communication-aided sensing*. On the one hand, communication systems can enhance sensing services by leveraging radio wave properties like transmission, reflection, and scattering to gain insights into the physical environment. Mobile communications networks also present significant opportunities for network sensing or sensing as a service. By sharing sensing results through the mobile network, multiple network nodes like base stations and terminals can collaboratively act as a sensing system. This collaborative sensing, achieved through sensing data fusion, reduces measurement uncertainty and expands the coverage area, enhancing sensing accuracy and resolution. Conversely, leveraging sensing information in communications becomes a key advantage of ISAC, as it results in a more predictable and deterministic propagation channel. This, in turn, enables the design and implementation of efficient communications algorithms and protocols, such as sensing-aided channel estimation, predictive beamforming with sensing assistance, rapid beam alignment and tracking, rapid recovery from beam failures, improved interference management, and link blockage mitigation, thus enhancing the QoS efficiency for communication systems. Furthermore, sensing effectively acts as a new channel, bridging the gap between the physical and digital worlds. Consequently, real-time sensing integrated with AI technologies plays a pivotal role in realizing digital twin, wherein virtual models faithfully mirror real-world entities and phenomena.

In the broader scope, the interaction between communication and sensing systems can be categorized as follows: (a) coexistence, where sensing and communication operate on separate hardware and may or may not share information, (b) cooperation, where the systems operate on distinct hardware but can share information to enhance each other, and (c) integrated design, where the systems are designed to function as a unified entity with shared information and joint design. ISAC's focus in future IMT centers around the integrated design approach. The evolution of ISAC technology can be divided into stages, ranging from loosely coupled to fully integrated. Beginning with resource sharing between communication and sensing systems, the evolution progresses to more closely intertwined stages where joint signal processing techniques and unified designs are employed. Ultimately, the mature stage of ISAC envisions full coordination and collaboration in various dimensions, leading to improved mutual performance, reduced costs, size, and power consumption.

The capabilities of ISAC open doors to a multitude of services that mobile operators can offer, including highly accurate positioning, tracking, imaging (e.g., for biomedical and security applications), simultaneous localization and mapping, environmental monitoring, gesture and activity recognition, and flaw/materials detection. These capabilities find applications in consumer and vertical contexts such as context-aware communications, industrial automation, connected vehicles, energy, and healthcare. However, challenges remain, including the need for shared hardware and waveforms, the fusion of information from different measurement sources, system-level design methodologies to assess trade-offs, solutions for heightened sensitivity to hardware imperfections, and the optimization of joint waveform designs.

12.12 Native AI

The success of AI in domains like image, video, audio signal processing, data mining, and knowledge discovery has paved the way for a shift toward an intelligent paradigm in wireless communication. This shift entails designing architectures, protocols, and algorithms for the IMT systems beyond 2030 by tapping into the untapped potential of wireless big data. The comprehensive utilization of this data remains largely unexplored. Consequently, with the extensive deployment of base stations, edge servers, and intelligent devices, mobile networks are poised to become a dynamic platform for seamless data collection, storage, exchange, and computation. These capabilities serve as potential catalysts for future mobile distributed and collaborative machine learning approaches. In the context of the forthcoming communication system, a revolutionary and transformative transition is on the horizon, enabling ubiquitous access to AI for individuals, businesses, and services anytime, anywhere. AI assumes the role of a design tool for shaping the future communication landscape, acting as the cornerstone for embedding intelligence universally. A key distinction between future communication systems and their IMT-2020 counterparts lies in their utilization of mobile technologies to foster the widespread adoption of AI and leveraging radio networks to amplify the presence of ubiquitous, distributed machine learning (ITU-R M.2516, 2022).

12.12.1 AI-Native Air Interface

The application of AI tools, such as deep learning, in the field of wireless communications has gained significant attention in recent years. This trend has largely been driven by the growing complexity of 5G systems and their evolution compared to previous generations of wireless technology. Deep neural networks have enabled the analysis of specific or unknown channel and network environments, encompassing factors like traffic, interferences, and user behaviors. This information is then used to adapt radio signaling to these conditions. Through learning, these networks can optimize user signaling, power usage, and end-to-end connectivity and even coordinate multi-user access to radio resources, thus enhancing data and control plane signaling and overall system performance.

The most formidable challenge in air interface design lies in understanding the communication environment, particularly in estimating and predicting fading channels (Jiang & Schotten, 2020). Traditional air interface approaches have dedicated considerable effort to pilot design and channel estimation. However, the advent of machine learning, especially the capabilities of deep neural networks for black-box modeling and hyper-parameterization, enables effective learning of the underlying channel characteristics, provided ample data is available. This acquired model can then be transferred to neighboring nodes using transfer learning, paving the way for novel air interface design strategies. Several components in the transceiver chain, including transmitter-side tasks like beamforming and management and receiver-side tasks such as channel estimation and symbol detection/decoding, are expected to be revamped using AI-based algorithms. As such, there will be a strong focus on redefining the physical layer of communication protocols using AI. Nevertheless, addressing the challenges related to the periodic updating of deep learning models used within various PHY layer blocks is of paramount importance.

Examples of areas ripe for investigation are detailed next:

- *AI in Signal Detection*: Machine learning techniques can play a pivotal role in symbol detection and decoding scenarios that deviate from the assumptions of optimal theory or where optimal solutions prove excessively complex. Given the trend toward shorter codewords and low-resolution hardware in upcoming IMT systems, which introduce challenging nonlinearities, machine learning can aid in symbol detection, precoding, beam selection, and antenna selection.
- *AI in Channel Estimation and Prediction*: Machine learning offers promise in estimating and predicting propagation channels (Jiang & Schotten, 2019), especially in scenarios with larger numbers of antenna elements, wider bandwidths, and high levels of time variation. Overcoming the limitations of previous CSI feedback schemes becomes feasible with AI-based approaches.
- *AI in MAC Layer Design*: The MAC layer stands as a prime application area for AI, where legacy problems can be replaced with data-driven AI methods involving supervised learning and model deployment. Future MAC algorithms need to account for coordination with AI functions across different layers of the network, particularly in the PHY layer.
- *AI in Radio Resource Management*: Radio resource management and allocation can be revolutionized by AI methods. Through reinforcement learning, base stations and user equipment can autonomously coordinate channel access and resource allocation based on received signals. This adaptive approach enhances interference coordination and system capacity.

- *Semantic Communications*: With advancements in machine learning and information theory, the ultimate air interface could achieve automatic semantic communications. However, challenges remain in obtaining wireless data for learning algorithms, which might be constrained by privacy issues. A solution involves learning from both practical wireless data and statistical models.

Crucial research topics in the near future include determining optimal machine learning algorithms under specific conditions, the required volume of training data, the transferability of parameters to different contexts, and enhancing explainability. AI's integration into radio interface technologies must be approached in phases to minimize disruptions in rollout and operation. In the short and medium terms, AI models can optimize specific features within IMT-2020 and its evolution. Over the long term, AI can introduce novel capabilities to legacy wireless systems.

12.12.2 AI-Native Radio Network

In the upcoming era of IMT systems, a paramount requirement is the provision of highly reliable and performance-assured services. These systems are set to introduce intricate multi-dimensional network topologies, thus amplifying the challenges of network management and operation. To counter these complexities, the integration of AI technologies for automated and intelligent networking services emerges as a feasible solution. Consequently, the radio access network of IMT systems beyond 2030 is projected to evolve into an AI-native architectural framework, primed to aid in computationally demanding tasks.

The pinnacle of an AI-native radio network involves the design and execution of intelligent network management (Jiang et al., 2017b), orchestrated by AI to dynamically optimize and adapt the network based on specific objectives, directives, or changes in the environment. Research in this domain encompasses high-level protocols, network architecture, and networking technologies that empower this intelligent radio network. Numerous scenarios enabled by AI-driven network automation (Jiang et al., 2019) have been proposed, encompassing fault recovery, root cause analysis, AI-fueled energy optimization, optimal scheduling, and network planning. However, the challenges related to training issues have been identified, such as performance limitations, lack of interpretability, uncertainties in generalization, and interoperability deficits required for full-fledged network automation. The classification of analytics into four types – descriptive, diagnostic, predictive, and prescriptive – opens avenues for future AI-native networks.

In the time to come, the RAN is anticipated to exhibit a heightened ability to perceive and adapt to complex and dynamic environments. This is achievable through continuous monitoring, tracking of conditions, diagnosis, and automated resolution of RAN issues. Realizing autonomy across the entire lifecycle management involves considering novel networking technologies, including:

- *Efficient and Intelligent Network Telemetry*: Leveraging AI to apply management operations based on a blend of historical and live network data
- *Automated Network Management and Orchestration*: Endeavoring to consistently seek the optimal state of the RAN and enforce appropriate management actions
- *Automated Lifecycle Management Operations*: Adjusting configurations in radio network elements, optimizing services and features during and after deployment
- *AI-Based Assistance*: Offering AI-driven support in aspects such as forecasting, root cause analysis, anomaly detection, and intent translation

Outlined below are examples of proposed research areas:

- *Intelligent Data Perception*: To circumvent data transfer burdens across network interfaces, intelligent data perception could be realized, utilizing techniques like Generative Adversarial Networks (GANs) to generate requisite data, thereby simulating real data and preserving data privacy. Establishing an open network data set and ecosystem could further advance this vision of zero-touch network management.
- *Incorporating User Feedback*: Introducing user feedback into network decision-making could enhance the understanding of user preferences and lead to more user-centric AI algorithmic decisions.
- *Pervasive Computation Nodes*: To support highly computation-intensive services, future IMT systems will necessitate widespread computation nodes across the network. This trend might involve redesigning control and user planes, as well as adopting technologies such as programmable switches and distributed/federated learning.

- *On-Demand Capability Supply*: An intelligent network that supplies on-demand capabilities is imperative for supporting diverse application scenarios, where AI no longer merely optimizes wireless resources, but integrates as a comprehensive intelligent system within the radio network.
- *Sensing and AI Collaboration*: The integration of new sensing and AI functions is pivotal to achieving network intelligence. This involves collecting, processing, and storing network data end-to-end through sensing, which can be accessed on-demand by AI functions to support various applications more efficiently.
- *Distributed and Unified AI Control*: Envisioning a distributed AI system with AI algorithms and models distributed across network functions, coordinated by a unified AI control center. This enables independent task execution, interaction, and measurement reporting within the distributed AI system.
- *Adaptive Solutions for Different Usage*: Tailored AI solutions are required to address distinct wireless domains, each with unique components, parameters, complexities, and temporal constraints. A variety of AI techniques can be employed to target these diverse problems, leading to potential benefits such as energy savings.

In conclusion, the imminent future of IMT systems mandates the integration of AI to foster reliable, adaptive, and intelligent network environments. These endeavors are poised to reshape the landscape of radio networks, forging a path toward unprecedented levels of efficiency and automation.

12.12.3 Network for AI-as-a-Service

The transition of the radio network is shifting from the over-the-top paradigm toward the realm of AI. Within wireless networks, the integration of AI applications and paradigms necessitates substantial data exchange, encompassing vast data volumes, machine learning models, and inference data shared among distinct entities in the network. Long-term platform technologies are a prerequisite to robustly support AI services, profoundly influencing the design of forthcoming radio networks, namely AI-powered radio networks. To optimally leverage computing and communication loads while adhering to local data governance and privacy regulations, the adoption of distributed and collaborative machine learning becomes crucial. Consequently, data-split and model-split strategies will be focal points in forthcoming research. These shifts profoundly impact future network design in three key ways:

- *Shift from Downlink-Centric to Uplink-Centric Radio*: Contrary to the current downlink-centric focus, where heavier traffic and superior QoS are directed toward downlinks, AI demands intensified model and data exchanges between base stations and the users they serve. This prompts a reevaluation of uplinks in network design to establish equilibrium, efficiency, and robustness in distributed machine learning.
- *Shift from Core Network to Deep Edge*: The geographical proximity of data and the computational and communication needs of deep machine learning pose significant challenges to end-to-end latency. To mitigate these challenges, novel networks and corresponding protocols must be devised. A potential avenue is the placement of predominant learning processes and threads in close proximity to the edge, thus forming a *deep edge* structure that effectively alleviates delays.
- *Shift from Cloudification to Machine Learning*: Given the decentralized nature of data and computing power, the communication and computation procedures inherent in machine learning algorithms span the entire network, spanning from cloud to edge and device. Therefore, the conventional approach of cloudification must also be rethought in terms of application-centricity, catering to the specific requirements of broadly distributed machine learning applications while strategically deploying computing and communication resources.

Furthermore, the future landscape of data-intensive, real-time applications necessitates the implementation of distributed AI solutions. These solutions are essential for enhancing human decision processes, constructing autonomous systems across various scales from small devices to complete factories, optimizing network performance, and managing the anticipated proliferation of billions of interconnected IoT devices. Given the inherent unreliability of heterogeneous IoT devices compared to high-performance centralized servers, the incorporation of distributed and self-organizing schemes is indispensable to fortify robustness against device and link failures. Presently, numerous unanswered questions persist in achieving the prerequisites of genuine distributed AI solutions. These encompass aspects such as data and resource distribution, distributed and online model training, as well as AI inference grounded in these models across diverse devices, locations, and domains characterized by varying degrees of context awareness. As for the future network architecture, it is anticipated to offer innate support for radio-based sensing and, through versatile connectivity, accommodate ultra-dense sensor and actuator networks. This configuration will enable hyper-local and real-time sensing and communication capabilities.

12.13 Summary

This chapter studies key technologies in radio transmission and networking that are able to realize the challenging performance for disruptive 6G use cases and applications. Through studying this chapter, you understand distinct categories of key 6G technologies – *New Spectrum* possibilities over THz and optical bands; *New Air Interface* advancements, including cell-free massive MIMO, ultra-massive MIMO, intelligent reflecting surfaces, and next-generation multiple access; *New Networking* prospects, including open radio access networks and non-terrestrial networks; and lastly, *New Paradigm*, driven by the convergence of communication, AI, and sensing.

12.14 Exercises

1. Describe the technical benefits of exploiting THz frequencies.
2. To realize THz communications and sensing, one of the most challenging issues is the small transmission distances of THz signals. What are the reasons for the short-range? How to deal with this issue?
3. During the design of the conventional cellular systems that operate in sub-6G frequencies, the engineers generally consider the channel effects like free-space path loss, shadowing, and multipath propagation. With the increase of carrier frequency, some other channel impairments become evident and should be seriously taken into account. Describe these channel impairments for THz frequencies.
4. The optical band can be divided into three different kinds of lightwave. Can you name them?
5. A wireless communication system relies on electronic components to generate and detect wireless signals. How does an optical wireless communication system work?
6. In a cell-free massive MIMO system, all data symbols intended for different users have to be spatially multiplexed so that all users' signals can be transmitted over the same time–frequency resource in the downlink. Which precoding methods are applied in cell-free massive MIMO to achieve this?
7. Identify the main difference between conjugate beamforming and zero-forcing precoding.
8. What are the particular advantages of intelligent reflecting surfaces?

Solutions for Exercises

Chapter 1

1. Key organizations include CEPT, 3GPP, 3GPP2, and IEEE. CEPT was initially responsible for the development of GSM, and 3GPP later took over the development and maintenance of GSM, as well as UMTS (which uses WCDMA), LTE, and 5G standards. 3GPP2 focuses on CDMA 2000 standards, and IEEE is involved in the standardization of WMAN technologies such as WiMAX.
2. Packet switching was introduced in 2G systems. It allowed for more efficient use of network resources and enabled data services like GPRS.
3. Key advancements in 2G include the transition from analog to digital transmission, the introduction of data services like SMS and GPRS, and the use of new modulation schemes.
4. The transition from 3G to 4G involved a shift from circuit-switched to packet-switched networks and the introduction of technologies such as MIMO and OFDM:
5. • 1G: Analog voice
   • 2G:Digital voice and SMS
   • 3G:Data services and mobile Internet
   • 4G:High speed data and IP-based services
   • 5G:Heterogeneous service types such as eMBB, mMTC, and URLLC
6. Anticipated features of THz communications, AI integration for network optimization, advanced sensing technologies, and the potential for integrating satellite and terrestrial networks for global coverage.

Chapter 2

1. In 1947, William R. Young at AT&T Bell Laboratories presented the cellular concept of the hexagonal geometry throughout a wide coverage area. Douglas H. Ring, also at Bell Labs, expanded on Young's initial concept. He sketched out the basic design for a standard cellular network and published the intellectual groundwork as a technical memorandum entitled *Mobile Telephony – Wide Area Coverage* in Bell Labs' internal journal on 11 December 1947. Per-cellular systems used a single high-power base station to cover an entire metropolitan area, where the frequency band was only used once without reuse. The small capacity of per-cellular systems cannot satisfy the rising demand for mobile telephone services, driving the advent of an elegant network design.
2. AMPS employed FDD to separate the downlink and uplink transmission, where the transmission from the mobile stations to base stations used the frequency band from 824 MHz to 849 MHz, while the 869 MHz to 894 MHz band was applied for the transmission from the base stations to mobile stations. In addition, AMPS adopted FDMA to divide the whole frequency band into a parallel of 30 kHz channels. With this spacing, 832 pairs of channels are available out of a 50 MHz frequency band.
3. The 1G cellular standards include U.S. AMPS, Scandinavia NMT, Japanese MCS, TACS from the United Kingdom, German C-450, and French Radiocom2000. All these standards offer analog voice services using frequency modulation. FDD and FDMA were the common selection for duplexing and multiple access.

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5

4. FDD employs two separate frequency bands for the uplink transmission and downlink transmission. FDD allows for simultaneous transmission and reception, making it ideal for real-time applications, such as phone calls or video conferences. But it needs paired frequency bands, which are sometimes hard to find. In contrast, TDD uses a single frequency band for both downlink and uplink but separates the transmission and reception of data in time. It is useful in data applications with an unbalanced amount of transmission and reception data. Furthermore, TDD is more flexible than FDD in terms of spectrum allocation.
5. The most critical component behind cellular network design is *frequency reuse* and *cell splitting*.
6. Cell splitting is used in cellular networks to ensure network scalability to handle the increasing demand for wireless services. During the initial phase of deploying a cellular network, it is efficient to cover an entire region or city with a few large cells. Partition a large cell into several smaller cells. In the process of splitting a cell, the power of the base station is decreased to cover a smaller area, and additional base stations are deployed to cover the original area. By reducing the size of each cell, the total number of available channels increases.
7. Sectorization is a technique employed in mobile communications to increase the capacity of a cellular system. By using directional antennas, the same frequency can be reused in different sectors, allowing more users without building new sites or network infrastructure.
8. When a mobile user moves from one cell to another cell, the communication quality decreases, or the connection is interrupted. To ensure uninterrupted service and guarantee the experience for mobile users, the connection needs to transition seamlessly between two cells.
9. A mobile system needs to simultaneously support a large number of active users, each of which requires the allocation of dedicated resources. To achieve this, orthogonal channelization techniques such as frequency-division, time-division, space-division, or hybrid combinations are employed to create dedicated channels.
10. FDMA divides system bandwidth along the frequency axis into multiple narrow-band channels. Each channel is allocated to a different user. Since each narrow-band channel suffers from frequency-flat fading, it does not require complex signal processing, making it a simple, cost-effective solution.

Chapter 3

1. Worse voice quality, limited system capacity, vulnerability to security attacks, limited international roaming, poor handover reliability, and expensive cell phones. These drawbacks were raised from the analog components and technologies used in 1G.
2. The transition from the first-generation to the second-generation cellular system was empowered by digital technology. A digital system can achieve a higher capacity than an analog system since digital communications can apply more spectral-efficient digital modulation and more efficient multiple access techniques. Digitization facilitates the compression of voice signals, the encryption of information against eavesdropping, and the support for data services. In addition, digital components are more powerful, more lightweight, smaller, cheaper, and more power-efficient than analog components.
3. Uplink – 890 MHz to 915 MHz and Downlink – 935 MHz to 960 MHz. DCS-1800 is a variant of the GSM standard operating on a higher frequency band, where 1710 MHz to 1785 MHz and 1805 MHz to 1880 MHz were assigned for the uplink and downlink transmissions, respectively.
4. B and C.
5. With the proliferation of Internet services in the late 1990s, the demand for mobile data services increased rapidly. 2G cellular technology was primarily designed for optimizing voice communication, with limited support for data services. The development of 2.5G cellular systems, as an enhancement of 2G, was motivated by the increasing demand for faster data transmission speeds and the need for novel mobile data applications.
6. Circuit-switched networks create a dedicated path for communication, reserving resources for the entire conversation duration. In contrast, packet-switched networks break data into packets, allowing flexible routing and shared resource usage. Circuit-switched networks retain a constant connection, even during pauses in communication, while packet-switched networks dynamically use resources as needed in a *best-effort* mode, adapting to changing conditions. Packet-switched networks are more scalable and efficient for diverse data loads, making them prevalent in Internet-oriented telecommunications and computer networks.
7. A wireless signal is distorted by interference, channel fading, and noise, which cause errors in the received data. Channel coding helps to protect the transmitted data against such errors by adding redundant information to the original data before transmission. It consists of two basic forms: forward error correction and error detection codes.

8. Every 20 ms speech signal is transformed into a transmission block with a length of 456 bits, resulting in a coding rate of 22.8 kbps.
9. You can apply discontinuous transmission. That is to suspend signal transmission and turn off the transmitter during the silence period. It has two major advantages: save power consumption and reduce mutual interference.

Chapter 4

1. 50 MHz (25 MHz in UL and DL each).
2. See sketch in Fig. 4.1. The MSs are the devices used by mobile service subscribers to access the services; the BSS is the GSM radio network that establishes efficient, reliable, and secure radio links between the MS and the fixed infrastructure; the NSS is the GSM backbone network that manages the call routing and switching between different MS and networks; and the OSS manages the operation and maintenance of the network, including monitoring, testing, and configuration management.
3. The IMEI is used to internationally identify MS, the IMSI is used to internationally identify subscribers, the TMSI is used to temporarily identify MS as a local replacement of IMSI (not internationally unique), while the LMSI is used as a searching key to enhance database operation efficiency. The separation of IMEI and IMEI is due to the separation of ME and SIM, which serves for the purposes of:
   (a) Decoupling the subscriber mobility from the equipment mobility, and
   (b) Allowing the subscriber to keep a consistent personalization of their service and data independently from the mobile terminal

   By using the TMSI instead of the acIMSI, the identity of the subscriber is completely hidden from the radio interface so as to protect privacy and enhance the GSM security. By using LMSI, the database operation efficiency is enhanced.
4. GMSK, FDMA/TDMA hybrid multi-access, and FDD/TDD hybrid duplex.
5. 200 kHz, about 4.615 ms, about 576.9 μs.
6. With frequency hopping.
7. RPE-LTP, coding rate = 244/260.
8. Cyclic code + convolutional code, coding rate = 112/121 (cyclic) ×121/228 (conv.) = 112/228 (overall).
9. $124 \times 8 - 1 = 991$. (NOTE: Only theoretical maximum for an isolated cell without neighbor. In practice it is much less than that.)
10. After power-up of the ME or re-inserting the SIM, the user will be required to enter the PIN to activate the SIM. After 3 consecutive attempts with incorrect PIN, the SIM will be blocked and can only be unblocked by a preset 8-digit PUK, which is also stored in the SIM. After 10 consecutive attempts with incorrect PUK, the SIM will be permanently blocked.
11. See Fig. 4.7.
12. See Fig. 4.9.
13. By the LU procedure – the LU process is executed:
    a. When the MS is registered with the PLMN, and
    b. When the MS enters a new LA
14. In GSM, the signal strength is measured not only in the current serving cell but also in its adjacent cells, and the handover decision is made with all adjacent cell measurements taken into account, as briefly illustrated in Fig. 4.14.

Chapter 5

1. During the 1990s, the boom of the Internet resulted in the proliferation of data services, such as email, web browsing, file sharing, search engines, interactive gaming, e-commerce, multimedia messaging, online high-fidelity music, and video streaming. Although packet-switching sub-networks were integrated, the 2.5G standards were not capable of providing sufficient data transmission rates and system capacity necessary to support Internet services. 3G aimed to create data-optimized systems to replace previous voice-centric cellular networks.
2. The phenomenal event in the era of 3G was the auctions of spectrum and licenses in Europe, which left many lessons for the mobile industry. The cost of 3G licenses varied widely across Europe, with some auctions raising significant amounts of money. This was a record-breaking amount at the time. The high expenditures on 3G licensing resulted in the relatively slow roll-out of 3G services in Europe.

3. The ITU Radiocommunication Sector took charge of specifying the minimal technical requirements of the 3G system. The related recommendation file is ITU-R M.1225.
4. C and D. The mainstreaming multiple access technology is CDMA, while WiMAX adopted a more advanced technique called OFDMA.
5. IMT-2000 CDMA direct spread was developed within 3GPP. This radio interface is called WCDMA, or the UMTS Terrestrial Radio Access (UTRA) FDD, where the term UTRA was renamed later as UMTS Terrestrial Radio Access, with the same abbreviation UTRA. Since GSM achieved a dominant role in the 2G market, WCDMA, as the evolution of GSM, inherited this advantage and also got a big success in the era of 3G.
6. Although WiMAX was developed late than other 3G standards, it adopted some pre-4G technologies like OFDM/OFDMA, MIMO, and LDPC codes.
7. The CDMA cellular network has the following features: Universal Frequency Reuse, Soft Capacity, Soft Handover, and Interference Sharing.
8. FDMA or TDMA assigns disjoint time–frequency slots to different users within the same cell and non-overlapping frequency bands to adjacent cells. In a CDMA system, the users not only in the same cell but also in different cells, including adjacent ones, share the same time–frequency resources. This results in a frequency-reuse factor of 1, referred to as *universal frequency reuse*.
9. The cost of soft capacity is the increasing level of mutual interference when more users participate in the communications.
10. The neighboring cells in narrow-band systems are assigned to different frequency blocks, and the mobile terminals can only tune to a single-carrier frequency at a time. Consequently, a terminal has to disconnect the outgoing cell before connecting to a new cell. Such a *hard handover* causes the drop of voice calls, which was one of the major problems worsening the user experience in the first- and second-generation systems. Since the neighboring cells use the same frequency in a CDMA system, a user at the edge of a cell can communicate with more than one base station simultaneously.

Chapter 6

1. It contains:
   - MSC, which handles call routing, registration, location updating, handovers, and SMS
   - SGSN, which handles all packet-switched data, such as user authentication
   - GGSN, which acts as an interface between the UMTS network and external packet-switched networks
   - HLR, which is a database used for storage and management of subscriber information
2. The UMTS CN is connected to UTRAN and GERAN via two parallel interfaces: the IuCS at between BSC and MSC for circuit-switched voice service, and the IuPS between RNC and SGSN for packet-switched data service, respectively. UMTS introduces the UMTS to replace HLR and AuC in GSM. Furthermore, UMTS supports IMS on top of its architecture.
3. NodeB is equivalent to a GSM BTS, and it connects to the user's mobile and handles the radio resources at the cell level. RNC manages the radio resources for one or more NodeBs and handles the handover decisions.
4. Having a separate control and user plane allows for more efficient use of resources. It enables better QoS handling and facilitates easier network upgrades.
5. Hard handover is a break-before-make process where the old radio link is released before a new one is established. In soft handover, the mobile stays connected to both the old and new cells during the handover process:
6. • Measurement reports are sent by the mobile device.
   - The decision to initiate a soft handover is made by the RNC.
   - New radio links are established while maintaining the old ones.
   - The old links are finally terminated, completing the handover.
7. • Commercial LCS for value-added services like navigation
   - Internal LCS for network optimization and planning
   - Emergent LCS for emergency services
   - Lawful Intercept LCS for legal tracking requirements
8. Emergency LCS is crucial for locating callers in life-threatening situations, aiding in timely and effective emergency response:

9. • CSCF: Call Session Control Function, which serves as the main signaling and control element
   • MGCF: Media Gateway Control Function, which handles the interconnection between IMS and legacy networks
   • MGW: Media Gateway, which enables the conversion of media streams between different network types

Chapter 7

1. 4G was developed with an end-to-end all-IP architecture where only a packet-switched network was provided to offer Internet-based services more effectively. This design completely abandoned the circuit-switched network, for the first time in the history of cellular systems, and voice service was transferred to a kind of data service known as VoIP.
2. LTE cannot entirely fulfill the requirements for IMT-Advanced. For instance, the expected peak data rates of 1 Gbps for low mobility and 100 Mbps for high mobility were not fully met. To close these gaps, an enhanced version in 3GPP Release 10 known as LTE-Advanced was recognized as one of the terrestrial radio interface technologies of IMT-Advanced.
3. LTE adopted OFDMA in the downlink, while SC-FDMA was in the uplink.
4. One of the main technical obstacles in OFDM transmission is the substantial fluctuations in the instantaneous power of the transmitted signals. It results in reduced efficiency and high cost of the power amplifier, imposing a specific design constraint for uplink transmission. SC-FDMA maintains a constant envelope with very low PAPR, which is appealing for mobile terminals, where low power consumption and low cost are required.
5. Actually, 30.72 MHz is the sampling frequency of LTE systems. The OFDM signals do not fully use all DFT points, where about 2048-1200=848 DFT points are just input zeros. Zero means turning off the corresponding subcarriers.
6. LTE-Advanced can support eight antennas to realize 8×8 MIMO.
7. Repeaters or *amplify-and-forward* relays just amplify the signals transparently, which does not affect the terminal or base station. Hence, it allows for their introduction into existing networks seamlessly without additional standardization. Bear in mind that not all technologies or components applied in a cellular system are standardized. For example, the receiver part is generally standard-free.
8. The performance gap between the center and edge in a cell is tremendous. The edge users suffer from not only weak received signals but also strong inter-cell interference. The aim of CoMP is to improve cell-edge performance.
9. To support a peak data rate of 1 Gbps, a large bandwidth is required. However, it is difficult to obtain large portions of the continuous spectrum due to intense competition for spectrum utilization and the fragmentation of legacy spectrum allocation. Moreover, the implementation of RF chains becomes hard with the increase in signal bandwidth. Therefore, several component carriers can be aggregated. LTE-advanced supports a maximum transmission bandwidth of 100 MHz using carrier aggregation.

Chapter 8

1. The main components of the LTE-A system architecture include the UE, eNodeBs, and the EPC, which consist of the MME, S-GW, P-GW, and other entities. The UE connects to the network via the eNodeB, which handle the radio communications. The EPC manages the user data sessions and mobility of the UE.
2. The architecture of E-UTRAN is more simplified than UTRAN. In E-UTRAN, the eNodeB takes on many of the responsibilities that were previously split between the NodeB and RNC in UTRAN, leading to a flatter architecture.
3. In transitioning from UMTS to LTE/LTE-A, the radio interface protocol stack underwent significant refinements. The BMC sublayer in UMTS was eliminated in LTE/LTE-A, with its functions absorbed by the RLC and PDCP sublayers. This streamlined the architecture, enhancing efficiency and flexibility. Concurrently, the PDCP sublayer was expanded to encompass both control and user planes, simplifying design, improving header compression, and consolidating encryption and integrity protection under one layer.
4. The PDCP layer in LTE-A handles IP header compression and encryption. The RLC layer manages the segmentation and reassembly of data packets. The MAC layer is responsible for scheduling and HARQ. The PHY layer deals with the actual transmission and reception of signals, utilizing advanced modulation schemes.
5. OFDMA is used in the downlink and allows multiple users to be served simultaneously by allocating them different subcarriers. SC-FDMA is used in the uplink and provides the benefits of OFDMA while maintaining a lower PAPR.
6. MIMO enhances the performance of LTE-A networks by using multiple antennas at both the transmitter and receiver ends. This allows for increased data rates and improved link reliability.

7. Carrier Aggregation in LTE-A allows for the combination of multiple carriers, potentially across different frequency bands, to increase the bandwidth and thereby the data rate.
8. Relaying in LTE-A involves the use of relay nodes to extend coverage, especially in areas where it is not cost-effective to deploy a full eNodeB. Heterogeneous Networks (HetNets) in LTE-A refer to the deployment of a mix of macro-, micro-, pico-, and femto-cells to improve both coverage and capacity.
9. The X2 interface in LTE-A connects eNodeB directly to each other. This direct communication facilitates faster handover decisions and coordination between eNodeB, especially in CoMP operations.
10. Inter-eNB handover in LTE-A is a handover that occurs between two eNodeB without the need to involve the core network, making the process faster and more efficient compared to traditional handover mechanisms.
11. The EPS-AKA mechanism in LTE-A provides mutual authentication between the UE and the network. It ensures that both the network and the UE can trust each other, and establishes encryption keys for secure communications.
12. NAS security in LTE-A ensures that signaling messages between the UE and the core network (MME) are encrypted and protected, preventing eavesdropping and tampering.

Chapter 9

1. Unlike previous cellular systems that focused on human-centric communication services, 5G aimed to broaden the scope of mobile communications. It extended beyond humans and includes objects, expanding from consumer applications to vertical industries. For the first time in the history of mobile communications, machine-type communications are centered as human-centric communication.
2. Three IMT-2020 usage scenarios include enhanced Mobile Broadband (eMBB), Ultra-reliable Low-Latency Communications (URLLC), and massive Machine-Type Communications (mMTC).
3. 5G needs to realize 20Gbps in the downlink and 10Gbps in the uplink.
4. The maximal bandwidths of 5G are differentiated according to the operating frequency bands – 100 MHz for sub-6GHz and 1 GHz for mmWave deployment.
5. All.
6. In an NSA deployment, the air interface is connected to the existing 4G EPC core network, allowing the capabilities offered by NR (such as lower latency) to be utilized without the need for network replacement. In contrast, SA deployment needs a 5GC core network.
7. NOMA is suitable for fulfilling the diverse requirements of 5G, like massive connectivity, high spectral efficiency, low latency, and fairness. It allows multiple users to reuse each orthogonal resource unit. However, superposition coding at the transmitter and successive interference cancellation at the receiver have been introduced, raising the hardware complexity.
8. SDN is used to separate the control plane from the data forwarding function.
9. NFV is used to decouple software from hardware.
10. LDPC leverages coding randomness by incorporating pseudo-random connections between variables and check nodes. LDPC codes have demonstrated excellent performance, leading to their successful adoption in WiMAX before 5G. In 2009, polar codes was introduced by Arikan, revolutionizing the construction of error-correcting codes and paving the way to achieve the Shannon capacity. The key concept behind polar codes is channel polarization.

Chapter 10

1. SA deployment is a 5G network setup that does not rely on the 4G LTE network infrastructure, whereas NSA deployment uses existing 4G infrastructure to support 5G data planes. NSA allows for a quicker and more cost-effective transition to 5G by leveraging existing investments.
2. Network slicing is a form of virtual network architecture using the same physical network. It allows the creation of multiple virtual networks that can cater to different service requirements, such as eMBB, URLLC, and mMTC:
3. • 5G NR uses a flexible waveform called CP-OFDM that is an evolution of the OFDM used in LTE-A. CP-OFDM is more adaptable to a wide range of frequency bands and deployment scenarios. It also supports a new frame structure that allows for dynamic TDD, which is crucial for optimizing network capacity and efficiency in different use cases.

- The radio frame design in 5G NR is more flexible compared to LTE-A, with scalable numerology (subcarrier spacing and symbol durations) that can be adjusted based on the deployment scenario. This flexibility allows 5G NR to efficiently serve everything from low-bandwidth IoT applications to high-throughput, and low-latency applications.
- The radio interface protocol stack in 5G NR has been optimized to reduce latency and improve efficiency. For instance, the RRC layer has been designed to support faster connection setup and release, the PDCP layer has been enhanced for more efficient header compression and encryption, and the SDAP layer has been introduced to map data flows to different QoS flows.

4. Architecture, see Fig. 10.9. CU handles the control plane and DU manages the user plane. RAN slicing allows operators to allocate resources dynamically to different virtual RANs, catering to specific service requirements.
5. The SBA of 5GC is characterized by the decoupling of NFs into standalone services that interact with each other via a common framework. This modularity allows for more flexible and scalable network deployments.
6. - In the 5G network, the Access and Mobility Management Function (AMFs) is primarily responsible for the management of all aspects related to mobility when the UE is in an idle or active state. It handles registration, deregistration, connection establishment, and release. The AMFs also play a pivotal role in reachability, while the UE is in idle mode and in the management of mobility when the UE moves between different access networks or within the same access network.
   - The SMF is responsible for session management, which includes establishing, maintaining, and terminating sessions. It also manages and stores session context related to the user plane.
   - Key mobility management procedures in 5G include Registration, which replaces the attach procedure in LTE; Deregistration, which replaces the detach procedure; Service Request, which establishes the user plane for data transfer; and Handover, which ensures service continuity when the UE moves between different gNodeBs. These procedures are designed to be more efficient and support the new service requirements of 5G, such as ultra-reliable low-latency communication.
7. - 5G introduces several new security protocols and features to address the challenges of user privacy and network integrity. The SEPP is introduced to protect inter-operator signaling. It ensures that user data and signaling information are securely exchanged between networks, using security mechanisms like topology hiding and application layer security.
   - UDM is responsible for user identification and supports authentication and authorization. It manages subscription data and oversees access authorization, which is crucial for maintaining user privacy and preventing unauthorized network access.
   - AUSF is a central part of the 5G security architecture, handling the authentication of users. It verifies the authenticity of the SIM/USIM credentials and ensures that only legitimate users gain access to the network. The AUSF works in conjunction with the UDM and the AMF to provide a robust authentication and security mechanism.
   - Together, these components form a comprehensive security framework that addresses various threats. SEPP secures the network edge, UDM manages subscriber data securely, and AUSF ensures robust authentication. This multi-layered approach is designed to protect against eavesdropping, man-in-the-middle attacks, and identity theft, among other security threats.

Chapter 11

1. Six usage scenarios of IMT-2030 include *Immersive Communication*, *Hyper-reliable and Low-Latency Communication*, *Massive Communication*, *Integrated Sensing and Communication*, *Ubiquitous Connectivity*, and *Integrated AI and Communication*.
2. EU's Hexa and Hexa-II projects, Finlish 6Genesis, German Open6GHub, etc.
3. Probably, terahertz technologies, integrated communications and sensing, AI-based communications, non-terrestrial networks, intelligent reflecting surface, and open RAN.
4. Digital twin, extensive reality, holographic communications, tactile Internet, Internet of Everything, AI, and ML. Some of these applications demand extreme data rates, e.g., 1 tera-bits-per-second, which is far beyond the peak rate of 5G. Some of these applications require near-zero latency.

5. For the usage scenario of *Integrated Sensing and Communication*, sensing-related capabilities will be defined by ITU-R. Also, AI-related capabilities will be specified for *Integrated AI and Communication*.
6. Providing communication services to sparsely populated areas is technically feasible but not economical. GEO communications satellites, although offering wide coverage, are costly to manufacture and launch. Moreover, the capacity of GEO is limited relative to its wide coverage.

Chapter 12

1. The THz band offers sufficient spectral resources for wireless communications, which is an effective solution for solving spectrum shortages in low-frequency bands. In addition, the tiny wavelength of THz waves provides high spatial resolution, offering high potential for THz sensing and positioning. THz also offers the particular capability of imaging. The smaller dimensions of THz enable nano-communications capabilities for nano-scale devices. THz is a potential band for realizing integrated communications and sensing.
2. The main causes of this problem include the high spreading loss that grows quadratically with carrier frequency, the gaseous absorption due to oxygen molecules and water vapor in the atmosphere, and the adverse effects of weather conditions. Such a propagation loss can reach hundreds of decibels per kilometer or even higher. High-gain directional antennas are necessary to compensate for such a high propagation loss. Due to tiny wavelengths, massive numbers of elements can be tightly packed in a small area to generate high beamforming gains.
3. (1) Atmospheric loss due to the absorption of oxygen molecules and water vapors; (2) weather effects such as rain and snow; and (3) blockage.
4. The optical band consists of infrared, visible light, and ultraviolet lightwaves.
5. An optical wireless communication system uses light-emitting diodes or laser diodes to emit optical signals at the transmitter, and then the receiver applies photonic diodes to detect the optical signals.
6. Cell-free massive MIMO applies conjugate beamforming or zero-forcing precoding to spatially multiplex the data symbols intended for all users in the downlink.
7. The philosophy of conjugate beamforming is to amplify the desired signal as much as possible while treating other users' signals as noise. In constant, zero-forcing precoding aims to completely cancel inter-user interference.
8. IRS can proactively adjust the propagation environment, which is the first time in wireless communications. In addition, IRS is passive, which implies low power consumption and low cost.

References

3GPP GERAN. (2000). Digital cellular telecommunications system (phase 2+); radio subsystem link control (3GPP TS 05.08 version 7.5.0 release 1998). Technical report, ETSI.

3GPP TR 38.801. (2017). Study on new radio access technology: Radio access architecture and interfaces (release 14), b Report TS 38.801 version 14.0.0 Release 14, The 3rd Generation Partnership Project.

3GPP TR38.913. (2020). Study on scenarios and requirements for next generation access technologies (Release 16), Report TR38.913, The 3rd Generation Partnership Project.

3GPP TS 22.261: Service Requirements for the 5G System; Stage 1; V16.16.0. (2021). Technical report, 3GPP TSG Services and System Aspects.

3GPP TS 23.060. (1999). General Packet Radio Service (GPRS): Service description, Specification TS 23.060, The 3rd Generation Partnership Project.

3GPP TS 23.501. (2021). System architecture for the 5G system (5GS); stage 2 (Release 17), Specification, The 3rd Generation Partnership Project.

3GPP TS 23.501. (2022). 5G; System architecture for the 5G System (5GS), Specification TS 23.501 version 15.13.0 Release 15, The 3rd Generation Partnership Project.

3GPP TS 23.501. (2023). 5G; Security architecture and procedures for 5G System, Specification TS 33.501 version 15.17.0 Release 15, The 3rd Generation Partnership Project.

3GPP TS 25.101. (2010). Universal Mobile Telecommunications System (UMTS); User Equipment (UE) radio transmission and reception (FDD), Specification TS 25.101 version 7.18.0 Release 7, The 3rd Generation Partnership Project.

3GPP TS 25.102. (2011). Universal Mobile Telecommunications System (UMTS); User Equipment (UE) radio transmission and reception (TDD), Specification TS 25.102 version 7.21.0 Release 7, The 3rd Generation Partnership Project.

3GPP TS 25.212. (2014). Universal Mobile Telecommunications System (UMTS); Multiplexing and channel coding (FDD), Specification TS 25.212 version 7.12.0 Release 7, The 3rd Generation Partnership Project.

3GPP TS 25.213. (2010). Universal Mobile Telecommunications System (UMTS); Spreading and modulation (FDD), Specification TS 25.213 version 7.7.0 Release 7, The 3rd Generation Partnership Project.

3GPP TS 25.214. (2010). Universal Mobile Telecommunications System (UMTS); Physical layer procedures (FDD), Specification TS 25.214 version 7.17.0 Release 7, The 3rd Generation Partnership Project.

3GPP TS 36.101. (2022). Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception, Specification TS 36.101 version 10.33.0 Release 10, The 3rd Generation Partnership Project.

3GPP TS 36.211. (2013). Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation, Specification TS 36.2111 version 10.7.0 Release 10, The 3rd Generation Partnership Project.

3GPP TS 38.101-1. (2023). 5G; NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone , Specification TS 38.101-1 version 15.22.0 Release 15, The 3rd Generation Partnership Project.

3GPP TS 38.212. (2022). 5G; NR; Multiplexing and channel coding, Specification TS 38.212 version 15.13.0 Release 15, The 3rd Generation Partnership Project.

3GPP TS 38.401. (2020). 5G; NG-RAN; Architecture description, Specification TS 38.401 version 15.9.0 Release 15, The 3rd Generation Partnership Project.

3GPP TS 45.002: GSM/EDGE multiplexing and multiple access on the radio path, V17.0.0. (2022), Technical report, 3GPP RAN.

6G Flagship Program. (2019). Key drivers and research challenges for 6G ubiquitous wireless intelligence, White paper, 6G Flagship, University of Oulu.

Akyildiz, I. F., & Jornet, J. M. (2016). Realizing ultra-massive MIMO (1024 × 1024) communication in the (0.06–10) Terahertz band. *Nano Communications Networks Journal, 8*, 46–54.

Akyildiz, I. F., Han, C., & Nie, S. (2018). Combating the distance problem in the millimeter wave and Terahertz frequency bands. *IEEE Communications Magazine, 56*(6), 102–108.

Alamouti, S. (1998). A simple transmit diversity technique for wireless communications. *IEEE Journal on Selected Areas in Communications, 16*(8), 1451–1458.

Alliance, N. (2015). 5G white paper.

Alliance, N. (2016a). Description of Network Slicing Concept, White Paper.

Alliance, N. (2016b). Recommendations for NGMN KPIs and requirements for 5G.

Andrews, J. G., Buzzi, S., Choi, W., Hanly, S. V., Lozano, A., Soong, A. C. K., & Zhang, J. C. (2014). What will 5G be? *IEEE Journal on Selected Areas in Communications, 32*(6), 1065–1082.

Arikan, E. (2009). Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Transactions on Information Theory, 55*(7), 3051–3073.

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5

Astely, D., Dahlman, E., Furuskär, A., Jading, Y., Lindström, M., & Parkvall, S. (2009). LTE: The evolution of mobile broadband. *IEEE Communications Magazine, 47*(4), 44–51.

Attar, R., Ghosh, D., Lott, C., Fan, M., Black, P., Rezaiifar, R., & Agashe, P. (2006). Evolution of cdma2000 cellular networks: Multicarrier EV-DO. *IEEE Communications Magazine, 44*(3), 46–53.

Baek, S., Kim, D., Tesanovic, M., & Agiwal, A. (2021). 3GPP New Radio Release 16: Evolution of 5G for industrial Internet of Things. *IEEE Communications Magazine, 59*(1), 41–47.

Berardinelli, G., de Temino, L. A. M. R., Frattasi, S., Rahman, M. I., & Mogensen, P. (2008). OFDMA vs. SC-FDMA: Performance comparison in local area IMT-A scenarios. *IEEE Wireless Communications, 15*(5), 64–72.

Berrou, C., & Glavieux, A. (1996). Near optimum error correcting coding and decoding: Turbo-codes. *IEEE Transactions on Communications, 44*(10), 1261–1271.

Berrou, C., Glavieux, A., & Thitimajshima, P. (1993). Near Shannon limit error-correcting coding and decoding: Turbo-codes (1). In *Proceedings of IEEE International Conference on Communications* (pp. 1064–1070).

Beyond 5G Promotion Consortium. (2022). Beyond 5G white paper: Message to the 2030s, White paper, Beyond 5G Promotion Consortium.

Bolcskei, H. (2006). MIMO-OFDM wireless systems: Basics, perspectives, and challenges. *IEEE Wireless Communications, 13*(4), 31–37.

Caire, G., & Shamai, S. (2003). On the achievable throughput of a multiantenna Gaussian broadcast channel. *IEEE Transactions on Information Theory, 49*(3), 1691–1706.

Chan, P. W. C., Lo, E. S., Wang, R. R., Au, E. K. S., Lau, V. K. N., Cheng, R. S., Mow, W. H., Murch, R. D., & Letaief, K. B. (2006). The evolution path of 4G networks: FDD or TDD? *IEEE Communications Magazine, 44*(12), 42–50.

Chang, R. W. (1966). Synthesis of band-limited orthogonal signals for multichannel data transmission. *Bell System Technology Journal, 45*, 1775–1796.

Chase, D. (1985). Code combining—a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets. *IEEE Transactions on Communications, 33*(5), 385–393.

Chaudhury, P., Mohr, W., & Onoe, S. (1999). The 3GPP proposal for IMT-2000. *IEEE Communications Magazine, 37*(12), 72–81.

Chen, S., Wang, Y., Ma, W., & Chen, J. (2012). Technical innovations promoting standard evolution: From TD-SCDMA to TD-LTE and beyond. *IEEE Wireless Communications, 19*(1), 60–66.

Chen, Y., Bayesteh, A., Wu, Y., Ren, B., Kang, S., Sun, S., Xiong, Q., Qian, C., Yu, B., Ding, Z., Wang, S., Han, S., Hou, X., Lin, H., Visoz, R., & Razavi, R. (2018). Toward the standardization of non-orthogonal multiple access for next generation wireless networks. *IEEE Communications Magazine, 56*(3), 19–27.

Chen, W., Montojo, J., Lee, J., Shafi, M., & Kim, Y. (2022). The standardization of 5G-Advanced in 3GPP. *IEEE Communications Magazine, 60*(11), 98–104.

Chih-Lin, I., Katti, S., Coletti, C., Diego, W., Duan, R., Ghassemzadeh, S., Gupta, D., Huang, J., Joshi, K., Matsukawa, R., Suciu, L., Sun, J., Sun, Q., Umesh, A., & Yan, K. (2018). O-RAN: Towards an Open and Smart RAN, White Paper, O-RAN Alliance.

China Mobile Communications Corporation, A., Huawei Technologies Co., Ltd, G., Deutsche Telekom AG, S., & Volkswagen, J. (2017). 5G service-guaranteed network slicing white paper. White Paper.

Choi, L.-U., & Murch, R. (2004). A transmit preprocessing technique for multiuser MIMO systems using a decomposition approach. *IEEE Transactions on Wireless Communications, 3*(1), 20–24.

Chung, S.-Y., Forney, G., Richardson, T., & Urbanke, R. (2001). On the design of low-density parity-check codes within 0.0045 dB of the Shannon limit. *IEEE Communications Letters, 5*(2), 58–60.

Clemm, A., Vega, M. T., Ravuri, H. K., Wauters, T., & Turck, F. D. (2020). Toward truly immersive holographic-type communication: Challenges and solutions. *IEEE Communications Magazine, 58*(1), 93–99.

Dahlman, E., Parkvall, S., & Sköld, J. (2011). *4G LTE/LTE-advanced for mobile broadband*. Academic Press, Elsevier.

Dahlman, E., Parkvall, S., & Skold, J. (2013). *4G: LTE/LTE-advanced for mobile broadband* Academic Press.

Dahlman, E., Parkvall, S., & Sköld, J. (2021). *5G NR—the next generation wireless access technology*. Academic Press, Elsevier.

Ding, Z., Lei, X., Karagiannidis, G. K., Schober, R., Yuan, J., & Bhargava, V. K. (2017). A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends. *IEEE Journal on Selected Areas in Communications, 35*(10), 2181–2195.

Donald, V. H. M. (1979). Advanced mobile phone service: The cellular concept. *The Bell System Technical Journal, 58*(1), 15–41.

Eberspächer, J., Vögel, H.-J., & Bettstetter, C. (2001). *GSM switching, services and protocols* (Vol. 2). Wiley Online Library.

Ehrlich, N. (1979). The advanced mobile phone service. *IEEE Communications Magazine, 17*(2), 9–16.

Elgala, H., Mesleh, R., & Haas, H. (2011). Indoor optical wireless communication: Potential and state-of-the-art. *IEEE Communications Magazine, 49*(9), 56–62.

Ericsson and Networks, N. (2014). Further LTE physical layer enhancements for MTC, Technical report, 3GPP TSG RAN Meeting #65.

Ericsson Report. (2020). Mobile data traffic outlook, Report, Ericsson.

Esmailzadeh, R., & Nakagawa, M. (2003). *TDD-CDMA for wireless communications*. Artech House.

Etemad, K. (2008). Overview of mobile WiMAX technology and evolution. *IEEE Communications Magazine, 46*(10), 31–40.

ETSI TC-SMG. (1996). Digital cellular telecommunications system (phase 2+); channel coding (GSM 05.03); version 5.2.0, Technical report, ETSI.

ETSI TR101178. (2001). A high level guide to the DECT standardization, Technical Report TR101178, European Telecommunications Standards Institute (ETSI).

EU Gigabit Connectivity. (2020). *Shaping Europe's digital future, Communication COM(2020)67*. European Commission.

Feng, H., Cui, Z., Han, C., Ning, J., & Yang, T. (2021). Bidirectional green promotion of 6G and AI: Architecture, solutions, and platform. *IEEE Network, 35*(6), 57–63.

Fettweis, G., Boche, H., Wiegand, T., Zielinski, E., Schotten, H. D., Merz, P., Hirche, S., Festag, A., Häffner, W., Meyer, M., Steinbach, E., Kraemer, R., Steinmetz, R., Hofmann, F., Eisert, P., Scholl, R., Ellinger, F., Weiß, E., & Riedel, I. (2014). The Tactile Internet, Technology Watch Report, ITU-T.

Fitzek, F. H. P., & Seeling, P. (2020). Why we should not talk about 6G. arXiv.

Fitzek, F., Boche, H., Stanczak, S., Gacanin, H., Fettweis, G., & Schotten, H. D. (2022). 6G activities in Germany. *IEEE Future Networks, 15*.

Foschini, G. (1996). Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. *Bell Labs Technical Journal, 1*(2), 41–59.
Foschini, G., & Gans, M. (1998). 'On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Personal Communications, 6*, 311–335.
Frenkiel, R., & Schwartz, M. (2010). Creating cellular: A history of the AMPS project (1971–1983). *IEEE Communications Magazine, 48*(9), 14–24.
Fukuda, E., Noda, A., & HiguchI, M. (2002). Overview of global standardization of IMT-2000 and its evolution. *Fujitsu Scientific and Technical Journal, 38*(2), 238–253.
Furuskar, A., Mazur, S., Muller, F., & Olofsson, H. (1999). EDGE: Enhanced data rates for GSM and TDMA/136 evolution. *IEEE Personal Communications, 6*(3), 56–66.
Gallager, R. (1962). Low-density parity-check codes. *IRE Transactions on Information Theory, 8*(1), 21–28.
Garg, V. (2010). *Wireless communications & networking*. Elsevier.
Gesbert, D., Shafi, M., Shan Shiu, D., Smith, P., & Naguib, A. (2003). From theory to practice: An overview of MIMO space-time coded wireless systems. *IEEE Journal on Selected Areas in Communications, 21*(3), 281–302.
Ghosh, A., Maeder, A., Baker, M., & Chandramouli, D. (2019). 5G evolution: A view on 5G cellular technology beyond 3GPP release 15. *IEEE Access, 7*, 127639–127651.
Goldsmith, A. (2005). *Wireless communications*. Cambridge University Press, Stanford University.
Han, S., Xie, T., & Chih-Lin, I. (2021). Greener physical layer technologies for 6G mobile communications. *IEEE Communications Magazine, 59*(4), 68–74.
Han, C., Wang, Y., Li, Y., Chen, Y., Abbasi, N. A., Kürner, T., & Molisch, A. F. (2022). Terahertz wireless channels: A holistic survey on measurement, modeling, and analysis. *IEEE Communications Surveys & Tutorials, 24*(3), 1670–1707. Third Quarter.
Heine, G., & Sagkob, H. (2003). *GPRS: Gateway to third generation mobile networks*. Artech House.
Holma, H., & Toskala, A. (2004). *WCDMA for UMTS-radio access for third generation mobile communications* (3rd ed.). Wiley.
Holma, H., & Toskala, A. (2006). *HSDPA/HSUPA for UMTS: High-speed radio access for mobile communications*. Wiley.
Holma, H., & Toskala, A. (2007). *WCDMA for UMTS: HSPA evolution and LTE*. Wiley.
Holma, H., & Toskala, A. (2011). *LTE for UMTS: Evolution to LTE-advanced*. Wiley.
Hu, Y., & Li, V. O. K. (2001). Satellite-based Internet: A tutorial. *IEEE Communications Magazine, 39*(3), 154–162.
Hu, J., Wang, Q., & Yang, K. (2020). Energy self-sustainability in full-spectrum 6G. *IEEE Wireless Communications, 28*(1), 104–111.
Huawei NetX2025. (2021). NetX2025 target network technical white paper, White Paper, Huawei.
Huawei VR Report. (2018). Cloud VR network solution white paper, White Paper, Huawei.
IMT-2030(6G) Promotion Group. (2021). 6G vision and candidate technologies, White paper, IMT-2030(6G) Promotion Group.
International Organization for Standardization. (1994). Information technology—Open Systems Interconnection – Basic Reference Model: The Basic Model, International Standard ISO/IEC 7498-1.
Irmer, R., Droste, H., Marsch, P., Grieger, M., Fettweis, G., Brueck, S., Mayer, H.-P., Thiele, L., & Jungnickel, V. (2011). Coordinated multipoint: Concepts, performance, and field trial results. *IEEE Communications Magazine, 49*(2), 102–111.
ITU-R M.1225. (1997). Guidelines for evaluation of radio transmission technologies for IMT-2000, Recommendation M.1225-0, ITU-R.
ITU-R M.1457. (2000). Detailed specifications of the terrestrial radio interfaces of International Mobile Telecommunications-2000 (IMT-2000), Recommendation M.1457-0, ITU-R.
ITU-R M.1645. (2003). Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000, Recommendation M.1645, ITU-R.
ITU-R M.2012. (2012). Detailed specifications of the terrestrial radio interfaces of International Mobile Telecommunications-Advanced (IMT-Advanced), Recommendation M.2012-0, ITU-R.
ITU-R M.2038. (2004). Technology trends, Report M.2038-0, ITU-R.
ITU-R M.2083. (2015). IMT Vision-Framework and overall objectives of the future development of IMT for 2020 and beyond, Recommendation M.2083-0, ITU-R.
ITU-R M.2134. (2008). Requirements related to technical performance for IMT-Advanced radio interface(s), Report M.2134, ITU-R.
ITU-R M.2150. (2022). Detailed specifications of the terrestrial radio interfaces of International Mobile Telecommunications-2020 (IMT-2020), Recommendation M.2150-0, ITU-R.
ITU-R M.2320. (2014). Future technology trends of terrestrial IMT systems, Report M.2320-0, ITU-R.
ITU-R M.2370. (2015). IMT traffic estimates for the years 2020 to 2030, Recommendation M.2370-0, ITU-R.
ITU-R M.2376. (2015). Technical feasibility of IMT in bands above 6 GHz, Report M.2376-0, ITU-R.
ITU-R M.2410. (2017). Minimum requirements related to technical performance for IMT-2020 radio interface(s), Recommendation M.2410-0, ITU-R.
ITU-R M.2411. (2017). Requirements, evaluation criteria and submission templates for the development of IMT-2020, Report M.2411-0, ITU-R.
ITU-R M.2412. (2017). Guidelines for evaluation of radio interface technologies for IMT-2020, Report M.2412-0, ITU-R.
ITU-R M.2516. (2022). Future technology trends of terrestrial international mobile telecommunications systems towards 2030 and beyond, Report M.2516-0, ITU-R.
ITU-R P.676. (2019). Attenuation by atmospheric gases and related effects, Recommendation P676-12, ITU-R.
ITU-R WP5D. (2020). Future technology trends for the evolution of IMT towards 2030 and beyond, Liaison Statement, ITU-R Working Party 5D.
ITU-T NET-2030. (2019). A blueprint of technology, applications and market drivers towards the year 2030 and beyond, White Paper, ITU-T Focus Group NET-2030.
Jiang, W., & Kaiser, T. (2016a). From OFDM to FBMC: Principles and comparisons. In F. L. Luo & C. Zhang (Eds.) *Signal processing for 5G: Algorithms and implementations*, chapter 3. Wiley/IEEE Press.
Jiang, W., & Kaiser, T. (2016b). From OFDM to FBMC: Principles and comparisons. In F. L. Luo & C. Zhang (Eds.) *Signal Processing for 5G: Algorithms and Implementations*, chapter 3. Wiley/IEEE Press.
Jiang, W., & Luo, F.-L. (2023). *6G Key technologies: A comprehensive guide*. IEEE Press/Wiley.

Jiang, W., & Schotten, H. (2019). Neural network-based fading channel prediction: A comprehensive overview. *IEEE Access, 7*, 118112–118124.
Jiang, W., & Schotten, H. D. (2020). Deep learning for fading channel prediction. *IEEE Open Journal of the Communications Society, 1*, 320–332.
Jiang, W., & Schotten, H. (2021a). Impact of channel aging on zero-forcing precoding in cell-free massive MIMO systems. *IEEE Communications Letters, 25*(9), 3114–3118.
Jiang, W., & Schotten, H. D. (2021b). Cell-free massive MIMO-OFDM transmission over frequency-selective fading channels. *IEEE Communications Letters, 25*(8), 2718–2722.
Jiang, W., & Schotten, H. (2022a). Deep learning-aided delay-tolerant zero-forcing precoding in cell-free massive MIMO. In *Proceedings of 2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall)*, London, UK.
Jiang, W., & Schotten, H. (2022b). Intelligent reflecting vehicle surface: A novel IRS paradigm for moving vehicular networks. In *Proceedings of 2022 IEEE 40th Military Communications Conference (MILCOM 2022)*, Rockville, MA, USA.
Jiang, W., & Schotten, H. D. (2022c). Initial access for millimeter-wave and terahertz communications with hybrid beamforming. In *Proceedings of 2022 IEEE International Communications Conference (ICC)*, Seoul, South Korea.
Jiang, W., & Schotten, H. D. (2022d). Initial beamforming for millimeter-wave and terahertz communications in 6G mobile systems. In *IEEE Conference on Wireless Communications and Networking (WCNC)*, Austin, USA.
Jiang, W., & Schotten, H. (2023a). Orthogonal and non-orthogonal multiple access for intelligent reflection surface in 6G systems. In *Proceedings of 2023 IEEE Wireless Communications and Networking Conference (WCNC)*, Glasgow, Scotland, UK.
Jiang, W., & Schotten, H. (2023b). Performance impact of channel aging and phase noise on intelligent reflecting surface. *IEEE Communications Letters, 27*(1), 347–351.
Jiang, W., & Schotten, H. D. (2023c). Cell-edge performance booster in 6G: Cell-free massive MIMO vs. reconfigurable intelligent surface. in *Proceedings of 2023 the 32nd European Conference on Networks and Communications (EuCNC) and the 6G Summit*, Gothenburg, Sweden.
Jiang, W., Cao, H., & Kaiser, T. (2014). Power optimal allocation in decode-and-forward opportunistic relaying. In *Proceedings of 2014 IEEE Wireless Communications and Networking Conference (WCNC)*, Istanbul, Turkey.
Jiang, W., Strufe, M., & Schotten, H. D. (2017a). Experimental results for Artificial Intelligence-based self-organized 5G networks. In *Proceedings of IEEE 28th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Montreal, Canada.
Jiang, W., Strufe, M., & Schotten, H. D. (2017b). Intelligent network management for 5G systems: The SELFNET approach. In *Proceedings of European Conference on Networks and Communications (EuCNC)*, Oulu, Finland (pp. 109–113).
Jiang, W., Strufe, M., & Schotten, H. D. (2018). An SDN/NFV proof-of-concept test-bed for machine learning-based network management. In *Proceedings of IEEE International Conference on Computing and Communications (ICCC)*, Chengdu, China.
Jiang, W., Anton, S. D., & Schotten, H. D. (2019). Intelligence slicing: A unified framework to integrate artificial intelligence into 5G networks. In *Proceedings of 2019 12th IFIP Wireless and Mobile Network Conference (WMNC)*, Paris, France (pp. 227–232).
Jiang, W., Han, B., Habibi, M. A., & Schotten, H. D. (2021). The road towards 6G: A comprehensive survey. *IEEE Open Journal on the Communications Society, 2*, 334–366.
Jorguseski, L., Pais, A., Gunnarsson, F., Centonza, A., & Willcock, C. (2014). Self-organizing networks in 3GPP: Standardization and future trends. *IEEE Communications Magazine, 52*(12), 28–34.
JPL. (1995). History of wireless communications. http://www.wirelesscommunication.nl/reference
Juarez, J. C., Dwivedi, A., Hammons, A. R., Jones, S. D., Weerackody, V., & Nichols, R. A. (2006). Free-space optical communications for next-generation military networks. *IEEE Communications Magazine, 44*(11), 46–51.
Kavitha, K., & Manikandan, S. (2015). LMMSE channel estimation algorithm based on channel autocorrelation minimization for LTE-advanced with adaptive guard interval. *Wireless Personal Communications, 81*(3), 1233–1241.
Kim, D., & Zarri, M. (2018). Road to 5G: Introduction and migration. White Paper.
Knisely, D., Kumar, S., Laha, S., & Nanda, S. (1998). Evolution of wireless data services: IS-95 to CDMA2000. *IEEE Communications Magazine, 36*(10), 140–149.
Kodama, M. (2002). The world's first 3G mobile phone service: A case study of innovation. *Journal of General Management, 28*(2), 5–13.
Larsson, E. G., Edfors, O., Tufvesson, F., & Marzetta, T. L. (2014). Massive MIMO for next generation wireless systems. *IEEE Communications Magazine, 52*(2), 186–195.
Letaief, K. B., Chen, W., Shi, Y., Zhang, J., & Zhang, Y.-J. A. (2019). The roadmap to 6G: AI empowered wireless networks. *IEEE Communications Magazine, 57*(8), 84–90.
Lin, Y.-B., Rao, H. C.-H., & Chlamtac, I. (2001). General packet radio service (GPRS): Architecture, interfaces, and deployment. *Wiley – Wireless Communications and Mobile Computing, 1*(1), 77–92.
Linge, N., & Sutton, A. (2014). The road to 4G. *The Journal of the Institute of Telecommunications Professionals, 8*(1), 10–16.
Liu, Y., Zhang, S., Mu, X., Ding, Z., Schober, R., Al-Dhahir, N., Hossain, E. and Shen, X. (2022). Evolution of NOMA toward next generation multiple access (NGMA) for 6G. *IEEE Journal on Selected Areas in Communications, 40*(4), 1037–1071.
Liu, J., Kato, N., Ma, J., & Kadowaki, N. (2015). Device-to-device communication in LTE-Advanced networks: A survey. *IEEE Communications Surveys & Tutorials, 17*(4), 1923–1940.
Liu, A., Huang, Z., Li, M., Wan, Y., Li, W., Han, T. X., Liu, C., Du, R., Tan, D. K. P., Lu, J., et al. (2022). A survey on fundamental limits of integrated sensing and communication. *IEEE Communications Surveys & Tutorials, 24*(2), 994–1034.
J. C. MacKay, D., & Neal, R. M. (1997). Near Shannon limit performance of low density parity check codes. *Electronics Letters, 33*(6), 457–458.
Markey, H. K., & Antheil, G. (1941). US2292387A: Secret communication system. US Patent.
Marzetta, T. L. (2010). Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Transactions on Wireless Communications, 9*(11), 3590–3600.
Marzetta, T. L. (2015). Massive MIMO: An introduction. *Bell Labs Technical Journal, 20*, 11–22.
Maxwell, J. C. (1865). A dynamical theory of the electromagnetic field. *Philosophical Transactions of the Royal Society of London, 155*, 459–512.
Melody, W. (2001). Spectrum auctions and efficient resource allocation: Learning from the 3G experiences in Europe. *Info, 3*(1), 5–13.
Mijumbi, R., Serrat, J., Gorricho, J.-L., Bouten, N., & Turck, F. D. (2016). Network function virtualization: State-of-the-art and research challenges. *IEEE Communications Surveys & Tutorials, 18*(1), 236–262.
Mishra, A. R. (2005). *Advanced cellular network planning and optimisation*. Wiley.

Mouly, M., & Pautet, M.-B. (1995). Current evolution of the GSM systems. *IEEE Personal Communications, 2*(5), 9–19.
NGMN. (2021). 6G drivers and vision (1st version), White paper, NGMN Alliance.
Ngo, H. Q., Ashikhmin, A., Yang, H., Larsson, E. G., & Marzetta, T. L. (2017). Cell-free massive MIMO versus small cells. *IEEE Transactions on Wireless Communications, 16*(3), 1834–1850.
Nunes, B. A. A., Mendonca, M., Nguyen, X.-N., Obraczka, K., & Turletti, T. (2014). A survey of software-defined networking: Past, present, and future of programmable networks. *IEEE Communications Surveys & Tutorials, 16*(3), 1617–1634.
Osseiran, A., Boccardi, F., Braun, V., Kusume, K., Marsch, P., Maternia, M., Queseth, O., Schellmann, M., Schotten, H., Taoka, H., Tullberg, H., Uusitalo, M. A., Timus, B., & Fallgren, M. (2014). Scenarios for 5G mobile and wireless communications: The vision of the METIS project. *IEEE Communications Magazine, 52*(5), 26–35.
Osseiran, A., Monserrat, J. F., & Marsch, P. (2016). *5G mobile and wireless communications technology*. Cambridge University Press.
Pareit, D., Lannoo, B., Moerman, I., & Demeester, P. (2012). The history of WiMAX: A complete survey of the evolution in certification and standardization for IEEE 802.16 and WiMAX. *IEEE Communications Surveys & Tutorials, 14*(4), 1183–1211.
Parikh, J., & Basu, A. (2011). LTE advanced: The 4G mobile broadband technology. *International Journal of Computer Applications, 13*(5), 17–21.
Peled, A., & Ruiz, A. (1980). Frequency domain data transmission using reduced computational complexity algorithms. In *Proceedings of 1980 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Denver, CO, USA (pp. 964–967).
Pickholtz, R., Schilling, D., & Milstein, L. (1982). Theory of spread-spectrum communications—a tutorial. *IEEE Transactions on Communications, 30*(5), 855–884.
Porter, P. (1985). Relationships for three-dimensional modeling of co-channel reuse. *IEEE Transactions on Vehicular Technology, 34*(2), 63–68.
Potter, A. (1992). Implementation of PCNs using DCS 1800. *IEEE Communications Magazine, 30*(12), 32–36.
Price, R., & Green, P. E. (1958). A communication technique for multipath channels. *Proceedings of the IRE, 46*, 555–570.
Qu, Z., Zhang, G., Cao, H., & Xie, J. (2017). LEO satellite constellation for Internet of Things. *IEEE Access, 5*, 18391–18401.
Qualcomm Incorporated. (2015). New work item: Narrowband IOT (NO-IoT), Technical report, 3GPP TSG RAN Meeting #69.
Rahnema, M. (1993). Overview of the GSM system and protocol architecture. *IEEE Communications Magazine, 31*(4), 92–100.
Rao, A. M., Weber, A., Gollamudi, S., & Soni, R. (2009). LTE and HSPA+: Revolutionary and evolutionary solutions for global mobile broadband. *Bell Labs Technical Journal, 13*(4), 7–34.
Rappaport, T. S., Sun, S., Mayzus, R., Zhao, H., Azar, Y., Wang, K., Wong, G. N., Schulz, J. K., Samimi, M., & Felix Gutierrez, J. (2013). Millimeter wave mobile communications for 5G cellular: It will work! *IEEE Access, 1*, 335–349.
Richardson, T., & Kudekar, S. (2018). Design of low-density parity check codes for 5G New Radio. *IEEE Communications Magazine, 56*(3), 28–34.
Richardson, T., Shokrollahi, M., & Urbanke, R. (2001). Design of capacity-approaching irregular low-density parity-check codes. *IEEE Transactions on Information Theory, 47*(2), 619–637.
Ring, D. H. (1947). *Mobile telephony—wide area coverage*. Bell Telephone Laboratories.
Rissen, J.-P., & Soni, R. (2009). The evolution to 4G systems. *Bell Labs Technical Journal, 13*(4), 1–5.
Sarieddeen, H., Saeed, N., Al-Naffouri, T. Y., & Alouini, M.-S. (2020). Next generation Terahertz communications: A rendezvous of sensing, imaging, and localization. *IEEE Communications Magazine, 58*(5), 69–75.
Schneider, P., & Urban, J. (2020). Die Sicherheitsarchitektur von Mobilfunknetzen. *White Paper, 1*, 4–7.
Schotten, H. D. (2023). Overview of German 6G program, Report, ETSI Research Conference 2023.
Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423.
Siles, G. A., Riera, J. M., & del Pino, P. G. (2015). Atmospheric attenuation in wireless communication systems at millimeter and THz frequencies. *IEEE Antennas and Propagation Magazine, 57*(1), 48–61.
Sollenberger, N., Seshadri, N., & Cox, R. (1999). The evolution of IS-136 TDMA for third-generation wireless services. *IEEE Personal Communications, 6*(3), 8–18.
Spencer, Q., Swindlehurst, A., & Haardt, M. (2004). Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels. *IEEE Transactions on Signal Processing, 52*(2), 461–471.
Steele, R., Lee, C.-C., & Gould, P. (2001). *GSM, CdmaOne and 3G systems*. Wiley.
Tariq, F., Khandaker, M. R., Wong, K.-K., Imran, M. A., Bennis, M., & Debbah, M. (2020). A speculative study on 6G. *IEEE Wireless Communications, 27*(4), 118–125.
Thalanany, S., Irizarry, M., & Saxena, N. (2017). License-assisted access considerations. *IEEE Communications Standards Magazine, 1*(2), 106–112.
Tse, D., & Viswanath, P. (2005). *Fundamentals of wireless communication*. Cambridge University Press.
Uusitalo, M. A., Rugeland, P., Boldi, M. R., Strinati, E. C., Demestichas, P., Ericson, M., Fettweis, G. P., Filippou, M. C., Gati, A., Hamon, M.-H., et al. (2021). 6G vision, value, use cases and technologies from European 6G flagship project Hexa-X. *IEEE Access, 9*, 160004–160020.
Vriendt, J. D., Laine, P., Lerouge, C., & Xu, X. (2002). Mobile network evolution: A revolution on the move. *IEEE Communications Magazine, 40*(4), 104–111.
Walke, B. H. (2003). The roots of GPRS: The first system for mobile packet-based global Internet access. *IEEE Wireless Communications, 20*(5), 12–23.
Walke, B. H. (2013). The roots of GPRS: The first system for mobile packet-based global internet access. *IEEE Wireless Communications, 20*(5), 12–23.
Watanabe, K., & Imamura, K. (1989). Evolution of NTT high-capacity land mobile communication system. In *Proceedings of 1989 IEEE International Conference on Communications (ICC)*, Boston, MA, USA.
Wong, D., & Lim, T. J. (1997). Soft handoffs in CDMA mobile systems. *IEEE Personal Communications, 4*(6), 6–17.
Wu, Q., & Zhang, R. (2020). Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network. *IEEE Communications Magazine, 58*(1), 106–112.
Wu, Y., Singh, S., Taleb, T., Roy, A., Dhillon, H. S., Kanagarathinam, M. R., & De, A. (2021a). *6G mobile wireless networks*. Springer.
Wu, Q., Zhang, S., Zheng, B., You, C., & Zhang, R. (2021b). Intelligent reflecting surface-aided wireless communications: A tutorial. *IEEE Transactions on Communications, 69*(5), 3313–3351.

Yan, L., Chen, Y., Han, C., & Yuan, J. (2022). Joint inter-path and intra-path multiplexing for terahertz widely-spaced multi-subarray hybrid beamforming systems. *IEEE Transactions on Communications, 70*(2), 1391–1406.

Yang, X. (2014). A multilevel soft frequency reuse technique for wireless communication systems. *IEEE Communications Letters, 18*(11), 1983–1986.

Yang, P., Xiao, Y., Xiao, M., & Li, S. (2019). 6G wireless communications: Vision and potential techniques. *IEEE Network, 33*(4), 70–75.

Young, W. R. (1979). Advanced mobile phone service: Introduction, background, and objectives. *Bell System Technical Journal, 58*(1), 1–14.

Yu, X., Xu, W., Leung, S.-H., & Wang, J. (2018). Unified performance analysis of transmit antenna selection with OSTBC and imperfect CSI over Nakagami-m fading channels. *IEEE Transactions on Vehicular Technology, 67*(1), 494–508.

Yuan, G., Zhang, X., Wang, W., & Yang, Y. (2010). Carrier aggregation for LTE-advanced mobile communication systems. *IEEE Communications Magazine, 48*(2), 88–93.

Yuan, Y., Yuan, Z., Yu, G., hwa Hwang, C., kai Liao, P., Li, A., & Takeda, K. (2015). Non-orthogonal transmission technology in LTE evolution. *IEEE Communications Magazine, 54*(7), 68–74.

Zaidi, A. A., Baldemair, R., Andersson, M., Faxér, S., Molés-Cases, V., & Wang, Z. (2017). Designing for the future: The 5G NR physical layer. *Ericsson Technology Review, 7*, 1–13.

Zhang, X. (2011). Analysis of M-sequence and Gold-sequence in CDMA system. In *2011 IEEE 3rd International Conference on Communication Software and Networks* (pp. 466–468).

Zhang, J., Bjornson, E., Matthaiou, M., Ng, D. W. K., Yang, H., & Love, D. J. (2020). Prospective multiple antenna technologies for beyond 5G. *IEEE Journal on Selected Areas in Communications, 38*(8), 1637–1660.

Zhou, X., Li, R., Chen, T., & Zhang, H. (2016). Network slicing as a service: Enabling enterprises' own software-defined cellular networks. *IEEE Communications Magazine, 54*(7), 146–153.

Index

W. Jiang, B. Han, *Cellular Communication Networks and Standards*, Textbooks in Telecommunication Engineering,
https://doi.org/10.1007/978-3-031-57820-5